

The Impact of Including Special Education Students In Accountability Systems

A Summary and Extension of a Paper Given at

The 31st Annual Conference on Large-Scale Assessment
Houston, TX
June 26,2001

Richard Hill

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

Background

For a variety of reasons, it is highly desirable to include all students, including special education students, in school scores to be used in accountability systems. One problem with doing this is that it is common to have some special education students, particularly those with moderate to severe mental disabilities, assessed in a completely different way from the majority of students. When this happens, Several issues arise regarding how the results for these students should be integrated into a total score for a school.

One approach is to attempt to put the results for all tests on a common scale. Numerous technical issues are associated with this approach. As a result, most discussions on this topic deal with potential solutions to the identified technical problems. This approach presumes that there is, in fact, a way of doing this: that the tasks these students perform measure the same constructs as the tasks given to others, but just at a different level of difficulty; if the proper equating is done, all the tasks can be put on one common scale. If the technical problems could be solved, there would be no further obstacles to combining the information.

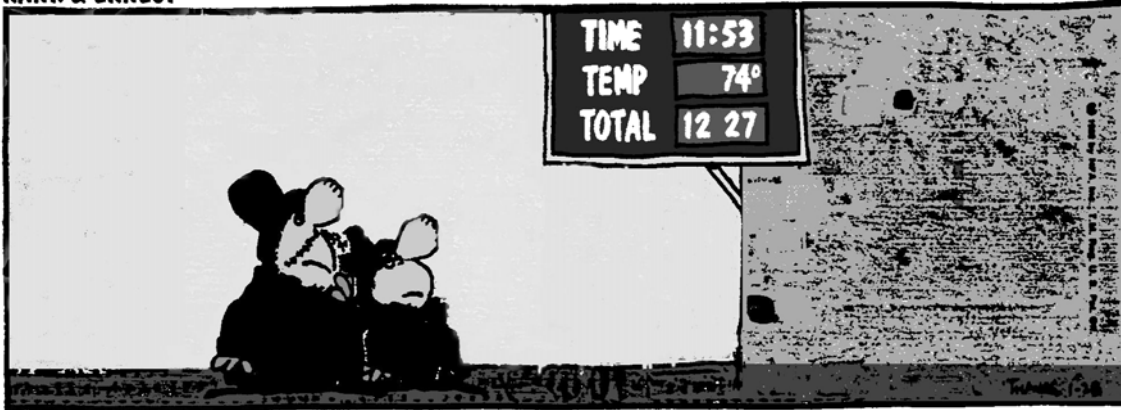
The problem with this approach is that the search seems destined to fail. The reason that no one has found a way to combine the information is not because we have insufficient technical skills, but because the tasks are truly different, and there is no technically correct way of placing the two tests on one scale.

This paper assumes that the technical problem cannot be solved, but that the search for a technical solution was pointless from the beginning. We will argue that the issue is one of policy, not technical correctness, and that within reasonable bounds, a wide variety of policies will yield similar solutions.

Scores vs. Gains

For years, this *Frank and Ernest* cartoon has been a favorite:

FRANK & ERNEST



The thought of combining “time” and “temperature” into a total may strike us as very funny; about as funny, to some people, as combining the reading scores of regular education students with the communication score for a student with a mental disability who is just learning to vocalize. Yet, our accountability systems are organized around “gain” and “improvement,” and we would feel that we could discern improvement in that student’s performance just as we could detect improvement in the reading skills of regular education students. So, as long as our interest is in detecting and rewarding improvement, the problem of combining scores could be rephrased in light of granting equal credit for equal amounts of improvement, rather than placing two different assessments on the same scale. As long as our *gain* scores made sense, it would not matter if the actual combination of scores in any one year was faulty.

That is, suppose there were two students on which I had data; one was a regular education student and one was in special education. Suppose further that I really could put their two very different test scores on a common scale of 0 to 100. Suppose that the achievement level of the regular education student on this scale is 85, and that of the special education student is 5. If correctly combined, the average for the two should be 45.

If two tests were scaled together incorrectly, the special education student might get a score of 25, rather than the score of 5 he should have. Now, the average of the two students would be 55, rather than the 45 it should be—a rather substantial error. This is the kind of example that people who have looked at this issue as a technical problem have been concerned about.

However, suppose that the intent is to evaluate people on their progress, not on their score. Suppose further that each student makes 5 points worth of progress for the year. The regular education student goes from 85 to 90, and the special education student goes from 5 to 10. If we had accurate assessment, we would observe that the average for the two students goes from 45 to 50, an average gain of 5 points.

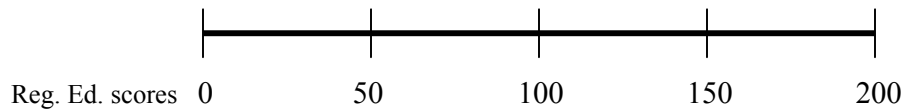
Now, suppose I evaluate my special education student on the faulty scale, and the student’s score increases from 25 to 30. The average for the two students would grow from 55 to 60. Neither of those two averages reflects the correct amount of achievement in the school, but the computation of a *gain* of 5 points IS correct. So long as our statistic of interest is gain, we have computed that statistic correctly despite the fact that the means for both individual years are inaccurate.

But how do we know whether the computation of gains is accurate? How do we know that the 5 points of gain shown by each student is an accurate reflection of their actual gain? If either of those numbers is incorrect, then perhaps we are right back where we started. The statement that the average gain is 5 is accurate only if we know that the amount of gain made by each student is the same. And if we can't place the original scores on the same scale, how can we be assured that the gains made on those two different scales are equivalent?

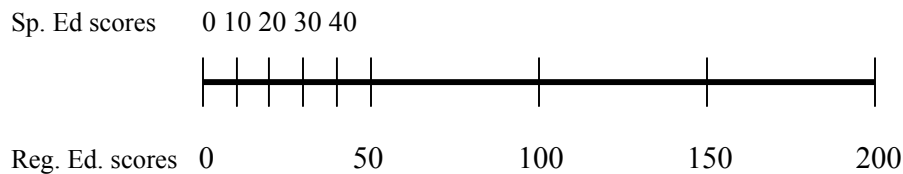
There are two facets to the answer to this question—neither of which directly addresses the issue of placing student scores on the same scale. First, we will show that, under certain conditions, gain scores are fairly robust to violations in scaling. Second, we will argue that the issue of how much to reward how much gain on each scale is largely a policy decision.

The Impact of Combining Incorrectly Scaled Scores

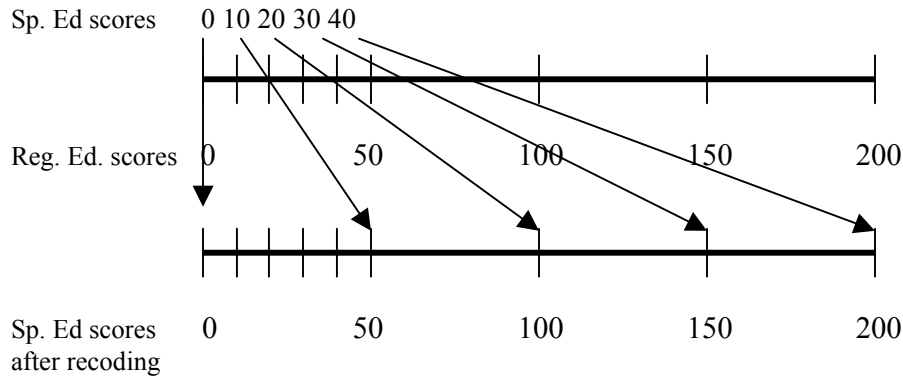
Suppose there is a statewide scale for achievement that places students into five different levels, and those levels are directly related to an underlying scale that ranges from 0 to 200. Such a scale could be represented as follows:



Suppose further that one could create a scale for disabled students that can, in fact, be correctly mapped onto the scale for regular education students. Suppose further that performance for the disabled students ranges from 0 to 40, in contrast to the 0 to 200 scale for the regular education students. This refinement could be represented as follows:



Now, suppose we decide to assign as many points for movement from one level to the next for special education students as we do for regular education students. Now, the scale for special education students will range from 0 to 200. This new scale is in error; each special education student's score is reported as being five times its value in the original scaling of the two tests.



We started by placing the special education students on the same scale as the regular education students, but changed that scale by multiplying every special education student's score by 5. While we might debate whether it ever was possible to put the special education scores and regular education scores on the same scale, there is no question that they are no longer on the same scale. Clearly, a score of 200 on the special education scale does not represent the same level of achievement as 200 on the regular education scale. Now, let's take a look at the consequences of what we've done.

Suppose, for purposes of our example, that the same proportion of regular education students is at each "level" as for the special education students (The issues—and final conclusions—would be the same for a wide range of examples. Making the two distributions the same simply makes the example easier to follow. Readers can construct several examples to satisfy themselves of the truth of the statement.) Let's further assume that both groups improve from Year 1 to Year 2, and that the amount of improvement (in terms of numbers of students moving from one level to the next) is the same for both groups. Some hypothetical data following those rules is provided in Table 1, along with the calculations for the school mean. Table 2 shows the same data as Table 1, but the scale for the special education students is now converted to the same 0-200 scale used for the regular education students.

Table 1

**Hypothetical Data for Two Years for Two Groups of Students
(Percentage of Students Falling Scoring at Each Assessment Level)**

Regular Education Students			Special Education Students		
"Correct" Scaled Score	Year 1	Year 2	"Correct" Scaled Score	Year 1	Year 2
200	10	10	40	10	10
150	20	25	30	20	25
100	30	35	20	30	35
50	25	20	10	25	20
0	15	10	0	15	10
Mean	92.5	102.5		18.5	20.5
Grand Mean Across All Students, If 90 Percent of Students Are Regular Education				85.1	94.3

Table 2

**Hypothetical Data for Two Years for Two Groups of Students
(Percentage of Students Falling Scoring at Each Assessment Level)**

Regular Education Students			Special Education Students		
“Correct” Scaled Score	Year 1	Year 2	Altered Scaled Score	Year 1	Year 2
200	10	10	200	10	10
150	20	25	150	20	25
100	30	35	100	30	35
50	25	20	50	25	20
0	15	10	0	15	10
Mean	92.5	102.5		92.5	102.5
Grand Mean Across All Students, If 90 Percent of Students Are Regular Education				92.5	102.5

If 10 percent of the scores are special education students, then the correct mean for Year 1 for the school is 85.1. If all the scores for the special education students are multiplied by 5, the mean for the school for Year 1 becomes 92.5. Even though they comprise only 10 percent of the school’s population, the alteration of their scale has changed the value of the school by over 7 points—an amount that would be considered excessive for many applications.

Now let’s take a look at the gains for the school. The correct score for Year 2 is 94.3—a gain of 9.2 points. The reported score, after the alteration of the scale, is 102.5, a gain of 10.0. That value is only 0.8 points higher than what the school would have received if the scale were accurate—an amount that is relatively small. Thus, while we inaccurately stated the actual performance of the school by quite a bit each year, the calculation of the *gain* made by the school was fairly close to accurate, even though we stretched the scale by a factor of 5 (which would be quite large compared to what most states propose doing) and the percentage of students in this category was 10 (which is a considerably larger percentage than most states are considering permitting in their “alternate assessment”).

However, suppose an additional change occurred between Year 1 and Year 2. Suppose all the regular education students who had scored 0 in Year 2 had now been given the alternate assessment; suppose further that they truly deserved scores in the range of 0 to 40 in the same proportion as the special education students. The actual achievement level of the school would be 94.3 (as in the above calculations—nothing has changed in the actual achievement levels of students in the second year, only the way we are counting their scores). But now, the reported score for the school would be 111.5, rather than 102.5. The reported gain would be over 17 points, when the school actually had gained 9.2. This example clearly illustrates the primary point of this section: *When measuring gain, it is far more important whether the conditions of testing and reporting are comparable across years than it is whether they are comparable within a year.* That is, the effect of using an incorrect scale is minimal so long as: (1) we evaluate schools on gain, not status, and (2) we keep the conditions of assessment (and the scale, however incorrect it might be) consistent from year to year.

Implications for Policy

Ultimately, the number of points awarded to anyone for improvement from Point A to Point B is a policy decision. The policy decision should be informed by technical considerations, but at the end of the day, policy makers will issue rules for the accountability system that are intended to reward those who accomplish the goals that the policy makers consider most important. There is no technical reason why achieving at the “Proficient” level in mathematics should be given equal credit to achievement at the same level in reading, or that improvement from “Basic” to “Proficient” is worth the same (or, in some systems, less than) improvement from “Proficient” to “Advanced.” All these are policy decisions, constrained to some minor degree by technical considerations.

Similarly, the appropriateness (or even the possibility) of scaling special education students’ scores on the same scale as that for regular education students could be debated endlessly. Even if we did reach a conclusion on that issue, we still would not have settled the issue in a way that would be helpful for policy makers. They need to answer the following questions: What rules are fair? What rules will encourage the greatest amount of improvement—for all students? What seems reasonable?

For example, in the illustrative data above, we supposed that we could convert the data for special education students to the same scale as that for the regular education students. We used the term “correct” to label that scale, meaning that if we attempted to measure the total educational output of the school, that scale would give us the correct answer. But on that scale, special education students get only 10 points for progressing from one level to the next, while regular education students get 50 points for a similar accomplishment. Policy makers might well judge that both achievements should be equally rewarded, and that a similar level of effort goes into accomplishing each. In that case, they would want to report results on the revised scale, regardless of its arguably deficient technical characteristics.

It is clear that there is no right or wrong answer to any of those questions. However, so long as policy makers stay within the parameters outlined in this paper—evaluate on gain, not status, and assure that testing conditions and practices remain constant from year to year—they should concentrate on selecting the answers to the above questions that they feel will best serve the needs of all populations of their state, and not worry about whether technical considerations will nullify the validity of their conclusions.

Advice for States Using a “Mixed Model”

While it is fine to point out that evaluations of gain are fairly robust to violations of accurate scaling, some states evaluate schools on both their status and their gain. With such designs, the state may have some sort of conjunctive rule—for example, that schools with high achievement have different requirements for improvement than schools with low achievement. In that case, the solution is not so obvious.

In such a situation, a school can alter its score considerably by miscoding special education students. If special education students are scored on the same 0-200 scale as the regular education students, for example, a school’s mean would go up considerably if it recoded some of its lower-scoring regular education students as being in special education. If schools vary in their implementation of the rules of which assessment to give to which student, there could be a considerable amount of error in school means. This, in turn, would place great pressure on schools that strictly follow the state’s guidelines on such coding to relax their implementation of those rules.

This situation, in turn, requires that the state ensure that rules are interpreted and implemented uniformly statewide. This can be accomplished through a series of actions, including:

- reviewing data on the number and percentages of students being reported as in special education (especially looking for changes over time)
- establishing limits on the percentage of students who can take alternative assessments
- establishing procedures for review of such decisions (such as requiring parent involvement in the decision)
- monitoring (and auditing, when results indicate such actions are warranted) schools' implementation procedures

The more a state relies on status, rather than gain, the more such actions are crucial to the correct interpretation of schools' results.