A Scale Drift Study

Jinghua Liu
Edward Curley
ETS, Princeton, NJ

Albert Low
National Board of Medical Examiners

Paper presented at the annual meeting of the
American Educational Research Association (AERA) and the
National Council on Measurement in Education (NCME)

April 13-17, 2009, San Diego, CA.
A Scale Drift Study

Scores on a test must be accurate and communicate meaningfully with the intended users. The score scale of a test provides a framework for score interpretation. Test-score interpretation can be facilitated by incorporating test contents, normative meaning and score precision into the score scale when it is established (Petersen, Kolen and Hoover, 1989). However, the score meaning may change for a variety of reasons: the norm group used at the time the scale was established may not be appropriate after a certain interval, and/or the test content may shift. Over time, the cumulative effects of these changes may be too large to be neglected. Another reason that score meaning may change is the use of equating. Scores are equated to provide score comparability. Nonetheless, equating is not perfect due to factors such as violations of equating assumptions and due to use of finite samples to estimate parameters (Haberman, Guo, Liu & Dorans, 2008). Hence, professional standards recommend that evidence be compiled periodically to document the stability of a score scale, if appropriate use of the scores depends on the stability of the scale (ETS, 2002).

Periodic monitoring of the SAT scale started in 1966. Stewart (1966) examined the extent of drift in the SAT-verbal (SAT-V) scale between 1944 and 1963. The results showed that SAT-V scores might have been 20 to 35 points higher in 1963 than in 1948. Modu and Stern (1975) assessed the SAT-V and SAT-math (SAT-M) scales between 1963 and 1973, and they found that the 1973 conversion was higher than the 1963 and 1966 conversions, in both the verbal and math sections. Both studies employed a non equivalent-group anchor test design. An anchor test from an old form was embedded in a
new form administration. Then the old form was placed on the new form scale through anchor test equating.

The last time a scale drift study was conducted for the SAT was in 1994 (McHale & Ninneman, 1994). The stability of the scale was assessed by placing the 1973-1974 forms and the 1974-1975 forms on the scale of 1983-1984 forms. Two equating designs were implemented. In the first design, a non equivalent-group anchor test design was used. Anchors from three 1973-1974 forms were embedded in the 1983-1984 administrations. In the second design, the operational sections of two 1974 and 1975 SAT forms were re-administered at 1984 administrations. The 1984 booklets contained one of the operational sections (v1, v2, m1 or m2) of the 1974 forms. Section pre-equating was conducted to obtain the new conversion of the 1974 forms. The results indicated that the SAT-V scale was relatively stable from 1973 to 1984, with little or no drift. However, the math results were inconsistent. The external anchor equatings indicated that the scale had drifted upward an average of 6 to 13 points, whereas the equatings based on re-administration of the old SAT forms indicated that the scale had drifted downward an average of 6 to 14 points. Overall, the previous studies all revealed that SAT scale drift did occur, at least within the 10-year interval, even though the results were inconsistent in terms of the direction of the scale drift.

The current SAT score scale was recentered nearly in 1995 and scores have been reported on the new scale since the May 1994 SAT administration (Dorans, 2002). It is critical that the scale stability is re-assessed to ensure the consistency of scale meaning.
Methodology

Equating Design

There are two types of data collection designs for equating employed in the SAT: the nonequivalent groups anchor test (NEAT) design and the equivalent groups (EG) design. At each SAT administration with one new form, the new form is equated to multiple old forms through a NEAT design. An EG design is usually employed in an SAT administration with two new forms, where the first new form is equated using the NEAT design and the second new form is equated to the first one through an EG design.

The spiraling procedure used in the SAT administration and the large numbers of test-takers who take each form usually ensure equivalent groups in the same administration.

For the purpose of current study, two tests were administered in the January 2005 SAT administration: a 1994 form and a more recent 2001 form. We treated the re-administered 1994 form as the new form, and we treated the re-administered 2001 form as the old form by applying the same conversion produced in 2001. We equated the re-administered 1994 form to the 2001 form via an EG design. It would have been desirable if the original 2001 form could have been directly linked back to the original 1994 form via the braiding plan so that the equating chain would be: re-administered 1994 form (in 2005) → 2001 form → originally administered 1994 form (in 1994). In this way, we could see the drifts more clearly, if there were any, since the test would be eventually equated back to itself. However, such a design was not possible due to the lack of availability of the SAT forms. Also, when we designed this study, our priority was to choose two forms that were as distant from one another as possible. Since the SAT was re-designed in 1994, which precluding the use of any test form prior to 1994, the
maximum possible distance from January 2005 was 11 years. Within this frame, the 1994 form and the 2001 form were the two most “distant” usable test forms.

**Content and Fairness Reviews**

Any time that a ten-year-old test is re-administered in the SAT program, it must be reviewed by ETS content experts to ensure that none of the questions has become outdated or inaccurate. In addition, it must be checked by trained fairness and sensitivity reviewers for alignment with the current ETS guidelines. Two different ETS fairness reviewers checked the 1994 test and found that no changes were required from the fairness perspective. However, verbal and math test development staff members also reviewed the 1994 test and, as a result, several changes were made before the test was re-administered in January 2005. In math, three questions were replaced because of changes between 1994 and 2005 in the particular calculators allowed for use during SAT administrations; in addition, a single word was added to the stem of one math question. In verbal, the date of publication (1985) was added to the introduction of the science passage, a footnote in the paired passage was reformatted to follow the current SAT style, and a minor wording change was made to the stem of one reading question. Thus, overall, 56 of the 60 math questions and 77 of the 78 verbal questions in this test form were identical in the 1994 and 2005 administrations.

**Equating Performed**

Since there were item revisions and item replacements in the re-administered 1994 form, statistical review was performed to determine whether the revised items performed similarly in the 2005 population as the original items performed in the 1994 population and whether a k-form equating is needed. A k-form test is created when slight
revisions are made to the original test. Such revisions may include dropping or replacing items, changing words, or adding additional information to an item. Consequently, a k-form equating is a process that links the newly revised form, the k-form, back to the original one via a single group design.

For the verbal section, the revisions were very minor. The revised items performed similarly across the 1994 and 2005 populations, with item characteristics such as difficulty (equated delta) and discrimination power (rbis) of the revised items being close to those of the original items (see Table 1). Therefore, it was decided that a k-form equating was not needed. The 2005 conversion was compared to the 1994 conversion directly.

For the math section, three items were replaced, and one item was revised slightly. The revision was minor, and the item characteristics were similar across the 1994 and 2005 populations. However, due to the fact that three items were replaced, the 2005 60-item conversion could not be compared to the 1994 60-item conversion directly. Instead, a 57-item k-form equating was conducted using the 1994 population to produce a 1994 k-form conversion; and another 57-item k-form equating was conducted using the 2005 population to produce a 2005 k-form conversion. The two k-form conversions were compared to each other.

**Equating Methods**

In this study, we performed mean-sigma and equipercentile equatings via the EG design to equate the 2001 form to the 1994 form to produce the new conversions for both verbal and math sections using the 2005 population. And then we also used mean-sigma and equipercentile methods to produce the 2005 k-form conversion for math using the
2005 population through single group design; and used mean-sigma and equipercentile methods to obtain the 1994 k-form conversion for math in the 1994 population. In this paper, we focus on the equipercentile equating only, because it was deemed the most appropriate for both the verbal and math equatings, and for the two k-form math equatings.

The equipercentile linking function is set so that the cumulative distribution function of scores on form $X$ converted to the form $Y$ scale is equal to the cumulative distribution function of scores on form $Y$ (Braun and Holland, 1982; Kolen & Brennan, 2004). This nonlinear transformation for the total population $P$ can be expressed as

$$e_Y(x) = G^{-1}[F(x)], \quad (1)$$

where $F$ represents the cumulative distribution function of $X$, $G$ is the cumulative distribution function of $Y$, and $G^{-1}$ is the inverse of the cumulative distribution function of $Y$. The intent is that the transformed scores on $X$ should have the same distribution function as the scores on $Y$, but the discreteness of the score scales will likely cause the cumulative distribution function for transformed scores on $X$ to differ somewhat from the cumulative distribution function for $Y$.

**Discrepancy Indices**

*Difference plots of conversions.* The difference plot, the new conversion minus the old conversion, is the most direct means of assessing score differences. At each score point level, the new conversion is compared to the old conversion.

To evaluate the relative magnitude of a difference in score conversions, Dorans and Feigenbaum (1994) proposed the notion of “score differences that matter” (DTM), in the context of SAT linking. On the SAT scales, scores are reported in 10-point units.
(200, 210, 220…780, 790, 800). For example, at a raw score of 53, the corresponding unrounded scaled scores might be 784.3 from new conversion and 785.1 from the old conversion. Due to the vagaries of rounding, the rounded reported scores would be 780 based on the new conversion, and 790 based on the old conversion, when ideally the rounded reported scores should be identical. The DTM, in contrast, treats these two conversions at this raw score point (53) as being equivalent. Dorans, Holland, Thayer and Tateneni (2003) adapted the above indices that is used in SAT practice, to other tests and considered DTM to be half of a score unit for unrounded scores. In the present study, the DTM was therefore defined as 5, which is half of the SAT score unit. Note this difference is best thought of as an indifference threshold. Any differences less than the DTM are considered not big enough to warrant any concern since they are smaller than the smallest difference that might actually matter.

**Root expected square difference (RESD).** The difference between the new conversion and the old conversion can be considered as conditional bias at each score point. We can also calculate the aggregated bias by averaging the conditional bias. However, the aggregated bias might be misleading since the differences with opposite directions can be offset with each other. Therefore, we calculate the root expected square difference (RESD) statistic instead, which is given by the formula

\[
RESD = \sqrt{\frac{1}{N} \sum f_x [s_{new}(x) - s_{old}(x)]^2},
\]

(2) to provide a standardized measure that evaluates the extent of the difference between the new conversion \(s_{new}\) and the original conversion \(s_{old}\) at each score level \(x\), where \(N\) is the total number of test-takers in the new form equating sample, and \(f_x\) is the frequency at
each score point. Note that we expressed the differences in raw-to-scaled score ($S$) units rather than in the raw-to-raw score units, in that most readers can understand and readily interpret the differences on the familiar College Board 200-to-800 scale.

**Percentage of score/examinees exceeding DTM.** In addition to using RESMD, we make use of the percentage of raw scores for which the new conversion differed from the old conversion by more than five points, and the percentage of examinees for whom these conversions create scores that differed by more than five points. The two stringent indices provide straightforward insights into lack of stability as a percentage of score range, and as a percentage of test-takers. The calculation of the two percentage indices is:

$$D_x = 1 \text{ if } |S_{\text{new}}(x) - S_{\text{old}}(x)| \geq DTM$$

$$\%\text{Affected Score Points} = \frac{\sum_x D_x}{X_{\text{max}} - X_{\text{min}} + 1}$$

(3)

$$\%\text{Affected Examinee} = \sum_x f_x D_x$$

**Results**

**Verbal Section**

The re-administered 1994 form was equated to the re-administered 2001 form in the January 2005 administration. The equating placed the 1994 form on the 2001 SAT scale. The new conversion was directly compared to the original conversion.

Figure 1 presents the differences between the original conversion in 1994 and the new conversion in 2005 (even though the new conversion is actually on the 2001 scale). When the dashed line is above the scaled score difference of 0, it represents an increase from the 1994 to the 2005 conversion. On the other hand, as the line drops below the
scaled score difference of 0 line, that represents a decrease from the 1994 to the 2005 conversion. As can be seen in Figure 1, the new conversion was higher than the old conversion across most of the score range. In other words, the 2005 test takers would have obtained slightly lower scores if the 1994 conversion were applied. Further, the difference curve fell within the DTM range (5 scaled score points) below scaled score of 300 and between score of 500-800. However, the differences were larger than 5 during the range of 300-500.

Table 2 summarizes the differences between the two conversions. For the 2005 test takers, means and standard deviations are listed when the 2005 conversion was used, and when the 1994 conversion was used. Also listed are the difference in means, the RESD value, the percentage of raw scores with an absolute unrounded scaled score difference equal to or larger than 5, and the percentage of examinees whose conversions resulted in scores that differ by at least 5 points. The data reveal that the 2005 test-takers had a higher mean with the 2005 conversion than they would have had with the 1994 conversion (7 points), which is not surprising given the new conversion moving upward as shown in Figure 1. The average RESD value is 6, which is larger than the DTM (5). The proportion of raw scores for which scaled scores between the two conversions differed more than 5 points was 31%. The percentage of examinees whose conversions resulted in scores that differed by more than 5 points was 53%. In summary, the results indicate an average drift upward of 6 scaled score points on the verbal section.

Math Section

Figure 2 displays the differences of two k-form math conversions, based on 57 common items, for the 1994 population and the 2005 population. As illustrated in Figure
2, the 2005 conversion line was below the 1994 conversion line across the entire score range. The differences were larger than DTM from 200 to 600, and fell within the DTM range from 600 to 800.

The differences are summarized in Table 2 as well. In general, the 2005 test takers obtained a lower mean with the 2005 conversion than they would had if the 1994 conversion had been used (6 points). The RESD value of approximately 6 was larger than the DTM of 5. The percentage of raw scores for which scaled scores between the two conversions differed more than 5 points was 55%. The percentage of examinees for whom the conversions differed by 5 or more was 73%, exhibiting a high degree of departure. In summary, the results indicated that the math scale has drifted downward 6 points on average, the opposite direction from verbal.

**Discussion**

This study examines the stability of the SAT scale from 1994 to 2001. A 1994 form was re-administered in a 2005 SAT administration, and was put on the 2001 scale by re-equating to a 2001 form. Both the verbal and math sections exhibit a similar degree of scale drift, but with opposite directions: the verbal scale has drifted upward, whereas the math scale has drifted downward.

There are a variety of reasons that a test scale may shift. First, the current SAT scale was established based on a 1990 Reference Group. This norm group may not be appropriate after more than 10 years. For example, the mean was set at 500 for both verbal and math when the SAT scale was recentered in 1995. Over the years, both verbal and math means have been progressively increasing. The means based on 2005 College-Bound Senior Cohort are 508 and 520 for verbal and math, respectively.
Second, test content and/or test takers’ familiarity with the content and item types may shift. In 1994, the SAT was revised and some major changes took place to both verbal and math sections. For verbal, the antonyms were removed, and the percentage of questions associated with passage-based reading increased. For math, two major changes took place: the introduction of student-produce-response (SPR) items, and the use of calculators. The SPR items had more exposure during the ten years since their introduction. As a result, the test takers in 2005 might perform better than the 1994 test takers on these items relative to other items. In addition, test takers in 2005 may be more familiar with using calculators in a test.

There are certain possible criticisms of these scale stability analyses that need to be pointed out. First, as discussed above, the re-administered 2001 form was not directly linked back to the 1994 form when it was equated in 2001. So we cannot evaluate the equating results based on an equate-back-to-itself chain. As a result, at least partial differences could be resulting from the form variations from test specifications. Second, the re-administered 1994 form was given in January 2005, while the original 1994 form was given in May. There are large ability differences, as a group, between the two administrations: May test takers have shown a higher ability than their January counterparts. This discrepancy between the two comparison groups may contaminate the equating results. Third, even though the revised items did not show significant differences in item characteristics, they may still have had some impact on the equating results. Finally, the accumulation of equating errors due to violation of equating model assumptions and due to use of definite samples can cause the instability of the scale as well.
Further investigation is needed to explain the opposite scale drift directions between verbal and math. One possible approach is to examine test takers’ performance by item types. For future research, we shall consider designing a braiding plan that will allow a systematic and periodic checking of scale stability.
References


Table 1

Comparison of Item Characteristics between the Original Items and the Revised Items

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Delta</td>
<td>rbis</td>
<td>Delta</td>
</tr>
<tr>
<td>Verbal</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>10.6</td>
<td>0.43</td>
<td>10.7</td>
</tr>
<tr>
<td>2</td>
<td>12.4</td>
<td>0.45</td>
<td>12.4</td>
</tr>
<tr>
<td>3</td>
<td>12.0</td>
<td>0.5</td>
<td>11.4</td>
</tr>
<tr>
<td>4</td>
<td>11.4</td>
<td>0.48</td>
<td>12.2</td>
</tr>
<tr>
<td>5</td>
<td>8.8</td>
<td>0.64</td>
<td>9.1</td>
</tr>
<tr>
<td>6</td>
<td>11.7</td>
<td>0.51</td>
<td>11.6</td>
</tr>
<tr>
<td>7</td>
<td>12.6</td>
<td>0.64</td>
<td>12.3</td>
</tr>
<tr>
<td>8</td>
<td>7.9</td>
<td>0.72</td>
<td>8.2</td>
</tr>
<tr>
<td>9</td>
<td>11.7</td>
<td>0.53</td>
<td>11.4</td>
</tr>
<tr>
<td>10</td>
<td>11.1</td>
<td>0.64</td>
<td>11.2</td>
</tr>
<tr>
<td>11</td>
<td>10.7</td>
<td>0.58</td>
<td>10.4</td>
</tr>
<tr>
<td>Math</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>15.5</td>
<td>0.65</td>
<td>14.6</td>
</tr>
</tbody>
</table>

Note. Item 2-11 in the verbal section is based on a reading passage, where the date of publication (1985) was added to the introduction of the science passage.
Table 2

*Summary statistics of scaled scores based on 2005 conversion and 1994 conversion in the January 2005 administration*

<table>
<thead>
<tr>
<th></th>
<th>Verbal</th>
<th>Math</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size</td>
<td>89,739</td>
<td>89,680</td>
</tr>
<tr>
<td>Mean &amp; SD based on 2005 conversion</td>
<td>487 108</td>
<td>496 111</td>
</tr>
<tr>
<td>Mean &amp; SD based on 1994 conversion</td>
<td>480 110</td>
<td>502 112</td>
</tr>
<tr>
<td>2005 conv mean - 1994 conv mean</td>
<td>7 -6</td>
<td></td>
</tr>
<tr>
<td>RESD</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>% FS with</td>
<td>unrounded scaled score diff</td>
<td>≥ 5</td>
</tr>
<tr>
<td>% Examinees with</td>
<td>unrounded scaled score diff</td>
<td>≥ 5</td>
</tr>
</tbody>
</table>
Figure 1. The verbal conversion differences between 2005 and 1994.
Figure 2. The math conversion differences between 2005 and 1994