Listening. Learning. Leading.®

Research Report
ETS RR–12-15

The Stability of the Score Scales for the *SAT Reasoning Test*™ From 2005 to 2010

Hongwen Guo

Jinghua Liu

Edward Curley

Neil Dorans

September 2012

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

The Stability of the Score Scales for the *SAT Reasoning Test*™ From 2005 to 2010

Hongwen Guo, Jinghua Liu, Edward Curley, and Neil Dorans
ETS, Princeton, New Jersey

September 2012

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Associate Editor: Rebecca Zwick

Technical Reviewers: Tim Moses and Gautam Puhan

Abstract

This study examines the stability of the *SAT Reasoning Test*™ score scales from 2005 to 2010. A 2005 old form (OF) was administered along with a 2010 new form (NF). A new conversion for OF was derived through direct equipercentile equating. A comparison of the newly derived and the original OF conversions showed that Critical Reading and Mathematics score scales have experienced, at most, a moderate upward scale drift (no greater than 5 points on average), and the drift may be explained by an accumulation of random equating errors. The Writing score scale has experienced a significant upward scale drift (11 points on average), which may be caused by sources other than random equating errors.

Key words: scale drift, test equating, equating error, *SAT*® test

Acknowledgments

The authors are grateful to Shelby Haberman for reviewing an initial draft of this report and also thankful to Rebecca Zwick, Tim Moses, and Gautam Puhan for reviewing the manuscript. Their comments and suggestions improve the quality of the paper. Any opinions expressed in this report are those of the authors and not those of ETS.

The *Standards for Educational and Psychological Testing* (American Psychological Association, Educational Research Association, & National Council on Measurement in Education, 1999, Part I, Chapter 4) require a testing program to check the stability of the score scales periodically. For a testing program that administers multiple test forms over successive years, equating is used to adjust test scores so that scores from different administrations are comparable. However, over time, test content evolution, change in the composition of the groups of examinees, and equating errors may result in *scale drift*, where this drift is defined to be a change in the interpretation that can be validly attached to scores on the score scale (Haberman & Dorans, 2011; Kolen & Brennan, 2004; Livingston, 2004). Even random error in equating can add up to a noticeable scale drift (Guo, 2010; Haberman, 2010). Adjustment of the scaling of the tests may be necessary for testing programs to maintain a meaningful scale. For example, the *SAT*® was rescaled in 1995 (Dorans, 2002), the ACT assessment was rescaled in 1989 (Brennan, 1989), and the Iowa Test of Basic Skills (ITBS) scale is adjusted every 7 years (Kolen & Brennan, 2004).

Periodic monitoring of the SAT-Verbal score scale has occurred for more than 50 years (Wilks, 1961). The drift in the SAT-Verbal scale before 1963 was assessed by Stewart (1966). It was found that the SAT-Verbal scale remained highly stable during the period from 1953 through 1963, but scale shift took place between 1948 and 1953.

Modu and Stern (1975) evaluated both the SAT-Verbal and SAT-Mathematics scales between 1963 and 1973. The equating procedures used in their study were identical to those used in the operational equating of SAT scores during the period of the study. The results indicated that the 1973 scale drifted upward by 14 points and 17 points on average for the SAT-Verbal and SAT-Mathematics tests, respectively.

McHale and Ninneman (1994) conducted an SAT scale drift study for the 1973–1974 forms and 1983–1984 forms. Two equating designs were implemented. One design was the nonequivalent groups with anchor test (NEAT), in which anchors from the 1973–1974 forms were embedded in the 1983–1984 forms. The second design involved the readministration of sections of 1974–1975 forms in the 1984 administrations. Overall, the results indicated that the score scale for SAT-Verbal was relatively stable from 1973 to 1984. However, the findings for SAT-Mathematics were inconsistent between the two equating designs: The NEAT design

equatings indicated an upward drift of the scale, while the equatings from the readministration of test forms indicated a downward drift of the scale.

Concerns about scale drift and possible score misinterpretation led the College Board and ETS to rescale the SAT, which was referred to as *recentering* (Dorans, 2002). The new scale was first used in the April 1995 administration.

Liu, Curley, and Low (2009) examined the SAT scale between 1994 and 2001 after the 1994 SAT redesign and before the 2005 SAT redesign. A 1994 SAT form and a 2001 SAT form were readministered in a 2005 administration. The 1994 form was equated to the 2001 form via the equivalent groups (EG) design using both linear and equipercentile methods. The results indicated that both the SAT-Verbal and SAT-Mathematics scales had drifted during 1994 to 2001, but in opposite directions: The SAT-Verbal scale moved upward, whereas the SAT-Mathematics scale moved downward.

There are other studies related to scale drift that are not specifically focused on the SAT. Petersen, Cook, and Stocking (1983) compared the extent to which different equating methods lead to scale drift. In their study, the linear, equipercentile, and item response theory (IRT) equating methods under the NEAT design were compared in a chain of equatings. It was found that the linear equating methods performed adequately for the type of equating situation used in their study.

Puhan (2009) studied the extent of scale drift on a test that employs cut scores. Equating results from both parallel equating chains and a single long chain were examined for three tests. The chained/direct linear equating and equipercentile equating methods were used to derive the conversions. Although some differences were observed in the conversions via different equating chains, the effect of these differences on the pass or fail status of test takers was not large. Some suggestions were also given in the paper on how to maintain scale stability: averaging conversions, selecting a short chain, and so on.

In this study, we investigate the scale stability of the SAT from 2005 to 2010. In 2005, the revised SAT was launched as the *SAT Reasoning Test*™ with three measures: Critical Reading (CR), Mathematics (M), and Writing (W) tests. The W measure was new, and the old SAT-Verbal was renamed as the CR measure. It is important to investigate whether the SAT Reasoning Test has been maintaining a stable scale for each of its three measures.

Method

The SAT Reasoning Test includes CR, M, and W measures. There are 67 multiple-choice items in CR, 44 multiple-choice items and 10 student-produced response items in M, and 49 multiple-choice items plus an essay in W.

Study Design

Several equating designs have been used in previous scale drift studies. We list three commonly used designs here. The first design involves the use of an external anchor test that is administered with both the old form (OF) and new form (NF), and then the OF is equated to the NF via the NEAT design to derive a new conversion for the OF. In this design, the external anchor is taken by both old and new groups and can be used to link the scores; the equating samples are drawn from the old administration and the new one, respectively. The second design is the spiraled sections design, where the sections of the OFs are spirally readministered with the NF. In the second design, items from the NF and sections of the OFs are calibrated so that scores on the new and OFs can be equated, usually via IRT equating. In the third design, an intact old test form is readministered with a NF. The NF is routinely equated, and the OF is equated to the NF using a direct equating method, such as the direct linear and/or equipercentile methods.

The first two data collection designs make restrictive assumptions. The first data collection design assumes that the anchor test adequately controls for differences in ability between groups over what is often a long period of time. The second data collection design assumes that researchers can piece together information about an entire test from information about small portions of the test (Holland & Wightman, 1982) or that IRT can be employed to do either true score equating or observed score equating. As Lord (1982) noted, these IRT methods require that each item measure the same dimension.

In contrast, the EG data collection design simply assumes that two forms are administered to groups that are equivalent. In this study, the third design was used (Figure 1). An OF from 2005 was readministered in a 2010 administration, along with the NF. The equating design of the NF followed the SAT operational test design—the NF shared common anchors with four OFs so that the NF was equated back to four OFs and an average conversion was obtained as the operational conversion line. The OF and NF test books were spiraled and distributed among test takers in such a way that it is expected that the group that took the NF is equivalent in ability to the group that took the OF.

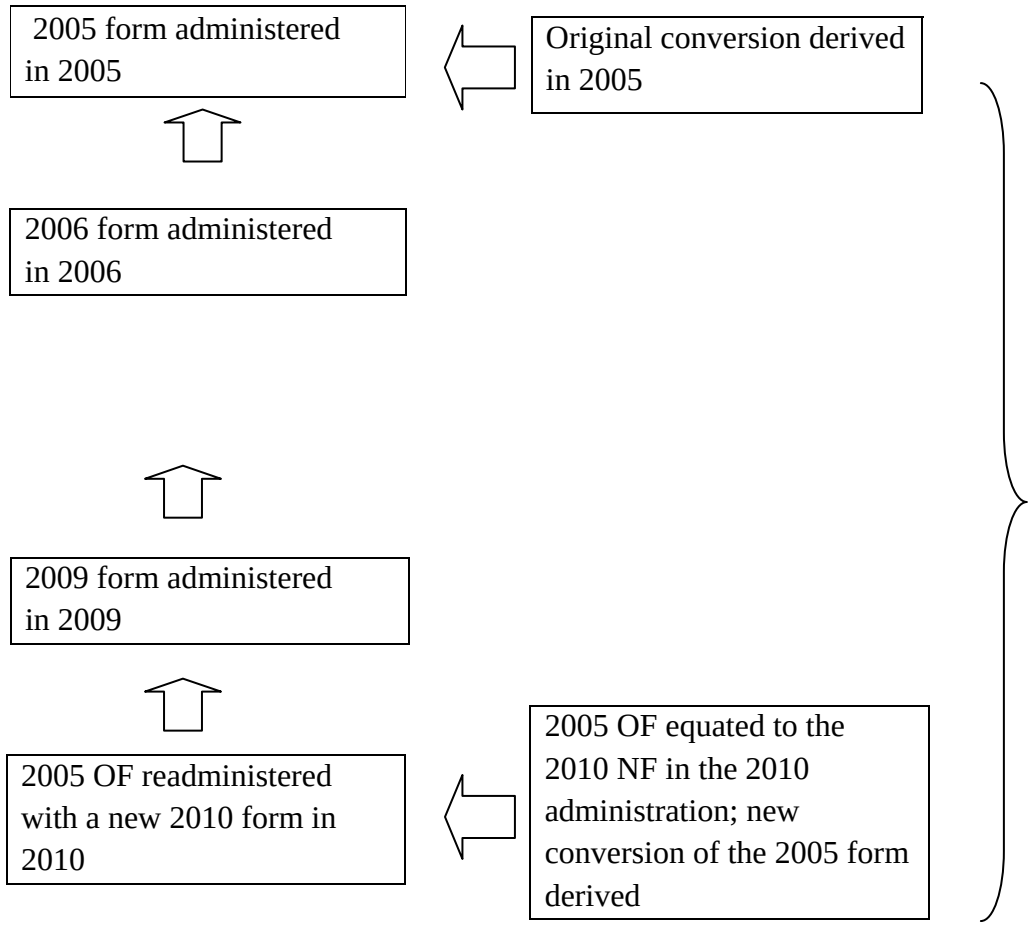

Figure 1. **The process of the current SAT Reasoning Test scale study design. NF = new form; OF = old form.**

Equating Methods

Data were collected to compare the abilities of the group that took the OF and the group that took the NF. Although the OF and the NF were spiraled in the same administration, an external anchor was still embedded in each form and administered with each form as a contingency plan. This is a typical design in SAT in case the spiraling procedure breaks down and the two groups are not equivalent. With an external anchor, there is still a mechanism to equate the two forms.

Equating samples were selected from juniors and seniors who took the SAT test in the administration. Table 1 displays the summary statistics of the two groups' raw scores on anchors

CR, M, and W. These sample sizes were above 6,000 for each of the three measures of the two forms. It was observed that the means and standard deviations were very similar. Therefore, the two groups were deemed equivalent, and the EG equating methods, both linear and equipercentile, were used to equate the OF to the NF to derive the new conversion for the OF for the scale drift study.

The equating samples were randomly drawn from the two groups who took the OF and NF forms in the 2010 administration. Table 2 provides summary statistics for CR, M, and W scores for the OF and NF. There were 120,000 test takers in each equating sample. The NF seemed to be easier in difficulty than the OF. Note that these sample sizes in Table 2 are much larger than those in Table 1. In this SAT administration, there are about 40 subforms, which are different in the variable section. Some variable sections are used as external anchors, and some are used as pretest sections. Each subform has roughly 6,000 test takers. In the administration, the NF and OF books were spiraled. In total, there were about 300,000 test takers. About half of the total test takers took the NF and half took the OF.

As in the operational equating, presmoothing was performed for both forms using a loglinear univariate model (Holland & Thayer, 2000), preserving six marginal moments. The direct equipercentile equating (i.e., the equipercentile equating under the random EG design) conversions were finally chosen as the operational conversions (see the Results section for details).

Table 1

Comparison of Raw Anchor Scores for the Two Groups (Junior & Senior Only) Taking OF and NF in 2010

	NF			OF		
Test	*N*	Mean	SD	*N*	Mean	SD
CR anchor	6,783	10.7	5.5	6,610	10.5	5.4
M anchor	6,769	9.2	4.7	6,658	9	4.7
W anchor	6,696	16.8	8.1	6,724	16.7	8.1

Note. NF = new form; OF = old form; CR = Critical Reading; M = Mathematics; W = Writing.

Table 2

Comparison of Raw Form Scores for the Two Groups (Junior & Senior Only) Taking Forms OF and NF in 2010

	NF				OF			
Test	*N*	Mean	SD	Skewness	*N*	Mean	SD	Skewness
CR	122,059	29	15.1	0.17	118,566	27.6	14.5	0.36
M		25.8	12.8	0.05		25.3	13.1	0.18
W MC		23	11.4	0.18		21	10.4	0.25

Note. NF = new form; OF = old form; CR = Critical Reading; M = Mathematics; W = Writing; MC = multiple choice.

Standard Error of Equating (SEE)

Because of random sampling in equating, equating error is inevitable. Standard errors of equating are used as indices of random error in equating (random equating error). For the direct equipercentile equating, the approximate random equating error of the raw score is

$$\text{var}[\hat{e}_Y(x_i)] \cong \sigma^2(Y)\frac{[P(x_i)/100][1-P(x_i)/100]}{\varphi^2}\left(\frac{1}{N_X}+\frac{1}{N_Y}\right), \quad (1)$$

where $\hat{e}_Y(x_i)$ is the estimation of the equated value of Y from x_i, $P(x_i)$ is the percentile rank (Kolen & Brennan, 2004, p. 249; Petersen, Kolen, & Hoover, 1989), and φ is the ordinate of the standard normal density at the unit-normal score, *z*, below which $P(x_i)/100$ of the cases fall. Lord (1982, pp. 166, 168) showed that $\hat{e}_Y(x_i)$ is asymptotically normally distributed. The SEE of the scale score for this study is calculated by using the method proposed in Kolen and Brennan (2004, p. 254). Notice that in this paper, we study whether the newly derived equating conversion drifted from the original conversion. Therefore we focus on the SEE of the NF equating in the following discussion and investigate whether the difference between the new and original OF scale score conversions are within the SEE band (that is, whether the difference can be attributed to the NF equating error).

Discrepancy Indices

To evaluate the relative magnitude of a difference in score conversions, we used the minimum differences that might matter (which is called *DTM for the SAT*; Dorans, Holland, Thayer, & Tateneni, 2003; Liu et al., 2009). On the SAT scale, scores for each of the three measures are reported in 10-point intervals (i.e., students' scores are reported as 200, 210, 220, …, 780, 790, 800). Hence there are two conversions each for CR, M, and W: the unrounded conversion obtained from equating and the rounded one in 10-point units. If two unrounded conversions differ by more than 5 points at a particular raw score, the rounded SAT score might be different. Therefore, a 5-point difference in conversions has some practical consequences.

The percentage of raw scores for which the new and original conversions differ by more than 5 points (% affected score points) and the percentage of examinees for whom the conversions differ by more than 5 points (% affected examinees) are calculated as follows:

$$\% \text{ affected score points} = \frac{\sum_x D_x}{X_{\max} - X_{\min} + 1}, \tag{2}$$

$$\% \text{ affected examinees} = \frac{\sum_x f_x D_x}{N}, \tag{3}$$

where D_x is 1 if the two conversions differ by at least 5 points at raw score x, otherwise D_x is 0; $X_{\min}$ and $X_{\max}$ are the minimum and maximum raw scores; f_x is the frequency of examinees at raw score x; and N is the number of total examinees.

Results

Table 3 summarizes the differences between scores obtained from the new equating conversions and scores obtained from the original equating conversions for the OF. These differences are discussed for CR, M, and W in the following subsections, respectively.

Critical Reading Scale

The OF was placed on the 200–800 scale, for the second time, by being equated to the NF via an EG design with 2010 data. As the relationship between the OF and NF was clearly curvilinear, the direct equipercentile equating was compared to the original OF conversion. The X-axis in Figure 2 represents the raw scores of the OF, and the Y-axis represents the difference

between the new and original OF scale score conversions (new – original, the solid line labeled *diff*) with the SEE band for the scaled difference: difference ±SEE (the two dashed lines). The SEE band is a pointwise range for the difference to fall in with a certain confidence level because of random sampling in equating (where L stands for the lower boundary and H for the higher boundary). The new equating conversion was significantly higher than the original for the bottom third of the score range (see Figure 2). For Raw Scores 65, 66, and 67, the reported score has been truncated to 800. Therefore, the differences at these top scores do not matter much. The biggest differences are in the lower end of the score range. For example, for raw scores below 4, the difference is larger than 20. However, the scaled scores are below 270 and the practical consequence of the differences may not be as important as they would be for scores in the middle range.

Table 3

Summary Statistics of Scaled Scores for the Old Form Based on the Original Conversion and Based on the New Conversion Derived in the Current Administration

	Critical Reading	Mathematics	Writing
Sample size	118,566	118,566	118,566
Mean and SD based on the new conversion derived by equating to the 2010 form using data from the 2010 administration	482 (105)	501 (111)	480 (107)
Mean and SD based on the original 2005 conversion	477 (109)	498 (113)	469 (108)
Mean difference (new - old)	5	3	11
RS with $\lvert$unrounded scaled score diff$\rvert \geq 5$	30	31	100
% Examinees with $\lvert$unrounded scaled score diff$\rvert \geq 5$	27	15	100

Note. RS = raw score.

As expected, for a sample size larger than 100,000, the SEE is relatively small (see Table A1 and Figure A1 in the appendix), indicating a high precision of equating. The SEE of scaled scores at the very top (Formula Scores 64 to 67) and the very bottom (Formula Scores -3 to 0)

are relatively large because of approximation and fewer numbers of examinees at those score levels.

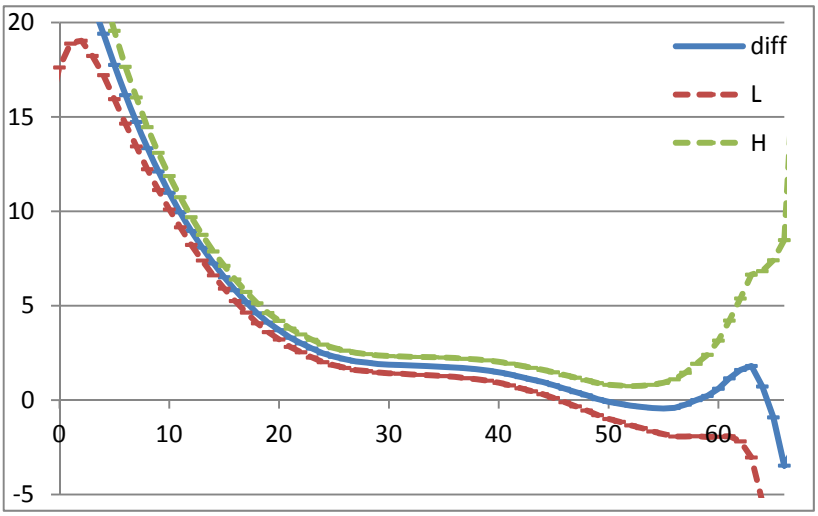

Figure 2. **The difference between the new and original CR raw-to-scale conversions and SEE, where L is the lower boundary and H is the higher boundary of the SEE band. CR = Critical Reading; SEE = standard error of equating.**

In addition, the scale score mean (482) produced by the new conversion was 5 rounded points higher than that produced by the original conversion, and this mean (482) was consistent with the NF mean (482). The proportion of raw scores for which scaled scores between the two conversions differed by at least 5 points was 30%. The percentage of examinees that would have been affected was 27%. Overall, the results indicate an average of 5.16 points upward drift on the Critical Reading section between the newly derived and the original OF conversions.

Mathematics Scale

The OF was placed on the 200–800 scale, for the second time, by being equated to the NF via an EG design with 2010 data. As the relationship between the OF and NF was clearly curvilinear, the direct equipercentile equating was compared to the original OF conversion. Figure 3 displays the differences between the new and original raw-to-scale conversions (the solid line labeled diff) with the scaled SEE band of the difference: difference ±SEE (the two dashed lines, where L is the lower boundary, and H is the higher boundary).

The new equating conversion was higher than the original across the majority of the score range (see Figure 3). For raw scores from 50 to 54 and from -3 to 9, the difference between the

two conversions is larger than 5. The practical consequence of the differences at the lower end of the score range may not be as important as for scores in the middle and upper ranges**.**

The SEE is relatively small, too (see Table A2 and Figure A2 in the appendix). The SEE of scaled scores at the very top (Formula Score 54) and the very bottom (Formula Scores -2 to -1) are relatively large because of approximation and fewer numbers of examinees at those score levels.

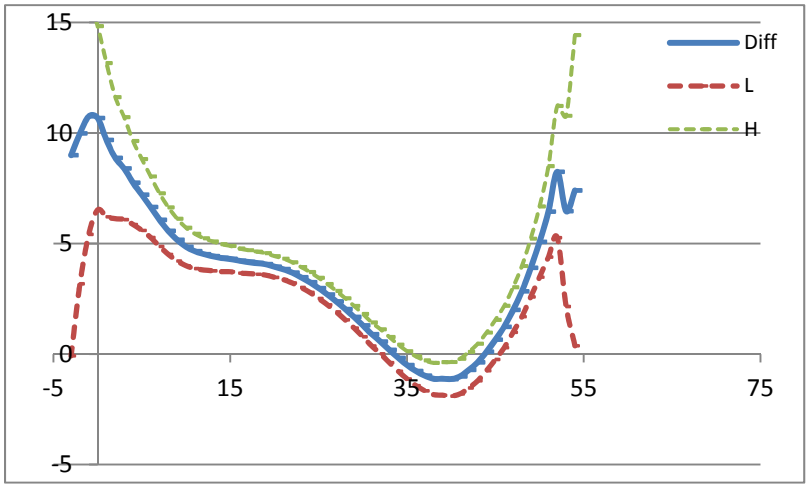

Figure 3. **The difference between the new and original Mathematics raw-to-scale conversions and SEE, where L is the lower boundary and H is the higher boundary of the SEE band. SEE = standard error of equating.**

In addition, the scale mean (501) produced by the new conversion was 3 rounded points higher than that produced by the original conversion, and this mean (501) was consistent with the NF mean (500). Therefore, it was appropriate to use the new direct equipercentile conversion as the operational conversion. The proportion of raw scores for which scaled scores between the two conversions differed at least 5 points was 31%. The percentage of examinees that would have been affected was 15%. Overall, the results indicate an average of 3.8 points upward drift on the Mathematics section between the newly derived and the original OF conversions.

Writing Scale

The OF was placed on the 200-to-800 scale, for the second time, by being equated to the NF via an EG design with 2010 data. As the relationship between the OF and NF was clearly curvilinear, the direct equipercentile equating was compared to the original OF conversion.

Figure 4 displays the differences between the new and original raw-to-scale conversions (the solid line labeled diff) with the scaled SEE band of the difference: difference ±SEE (the two dashed lines, where L is the lower boundary, and H is the higher boundary). The new equating conversion was significantly higher than the original across the entire score range (see Figure 4). Scaled scores at Raw Scores 48 and 49 for W were not obtained with this new equating, so they are omitted from Figure 4. For the whole score range, the difference between the two conversions is larger than 5. The practical consequence of the differences at the lower end of the score range may not be as important as for scores in the middle and upper ranges.

Table A3 and Figure A3 in the appendix display the SEE of the equating. The SEE of scaled scores at the very top (Formula Scores 48 to 49) and the very bottom (Formula Scores -3 to 0) are relatively large.

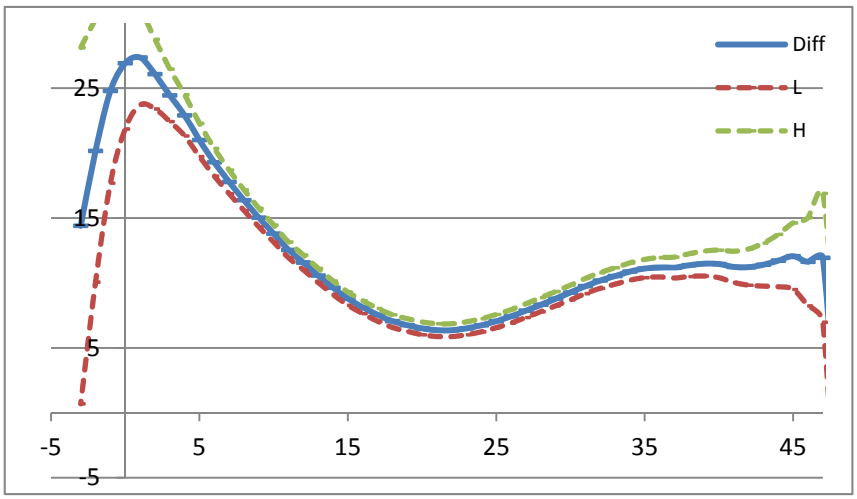

Figure 4. **The difference between the new and original Writing raw-to-scale conversions and SEE, where L is the lower boundary and H is the higher boundary of the SEE band. SEE = standard error of equating.**

In addition, the scale mean (480) produced by the new conversion was 11 rounded points higher than that for the original conversion, and this mean (480) was consistent with the NF mean (480). Therefore, it was appropriate to use the new direct equipercentile conversion as the operational conversion. The proportion of raw scores for which scaled scores between the two conversions differed by at least 5 points was 100%. The percentage of examinees that would

have been affected was also 100%. Overall, the results indicate an average of 11.20 points upward drift on the W measure between the newly derived and the original OF conversions.

Discussion

From the above analysis, it is evident that the SAT scales have drifted upward for the three measures of the OF forms in the past 5 years, especially for the W measure. There are a variety of reasons for scale drift, including test construction practices, subpopulation shifts and changes in populations, sampling errors, accumulation of random equating error, anchor design, and model misfit (Haberman & Dorans, 2011).

What could cause the drift of the SAT scales over the years? Here we calculate the equating error caused by random sampling and measure whether the drift in the SAT scales can be explained by the random sampling error. Random equating error can also accumulate after a series of equatings. In addition to the SEE in the EG equating from OF to NF, the cumulative SEE (ASEE) of regular SAT equatings under the NEAT design will add up. Assuming all the SEEs in each single SAT equating are very similar, the ASEE can be estimated using a rule of thumb (Guo, 2010) in which $\sqrt{m}$ is multiplied by the SEE in a single equating, where m is the number of equatings from the NF form back to a 2005 form. In this study, $m = 5$. To estimate the SEE in one regular SAT equating, we use (1) again for simplicity. The estimated SEE from (1) is likely to be less than the SEE in the operational SAT equatings. Notice that the regular equating sample size is usually around 6,000. The approximate total ASEE of the OF form equating from 2005 to 2010 for CR, M, and W is displayed in Figures A1, A2, and A3, respectively.

In Figures 5–7, the solid line (diff) is the unrounded difference between the original and newly derived OF conversions and the dashed lines (CL and CH) are the band of the estimated ASEE. For CR, it can be observed in Figure 5 that the zero line is within the band of the ASEE except for a few raw score points around 10.

For M, it can be observed in Figure 6 that the whole zero line is within the ASEE band. In some sense, these observations may indicate that the scale drift of CR and M is within the expected range—the accumulated random equating error alone can account for the observed amount of scale drift.

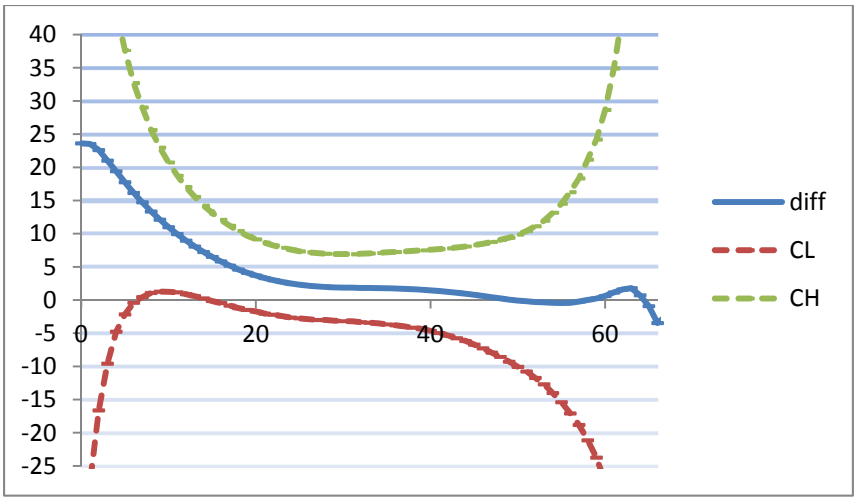

Figure 5. **The difference of the new and original raw-to-scale conversions of OF with the accumulative SEE for CR, where CL is the lower boundary and CH is the higher boundary of the SEE band. CR = Critical Reading; OF = old form; SEE = standard error of equating.**

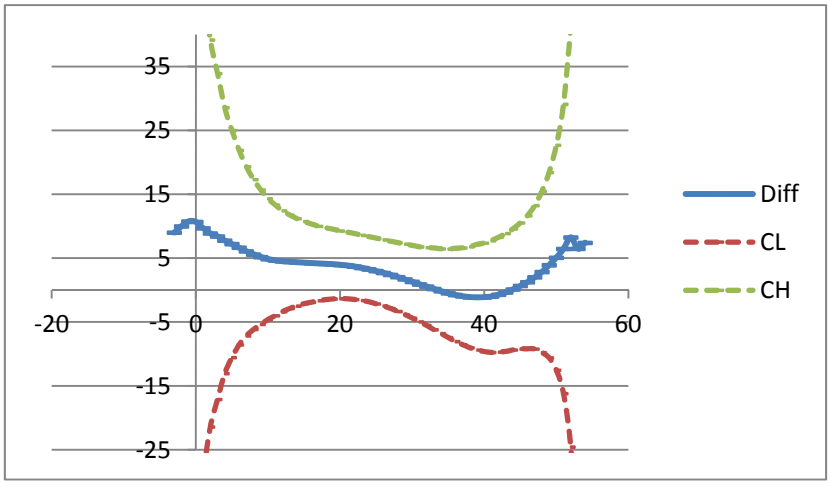

Figure 6. **The difference of the new and original raw-to-scale conversions of OF with the accumulative SEE for M, where CL is the lower boundary and CH is the higher boundary of the SEE band. M = Mathematics; OF = old form; SEE = standard error of equating.**

However, for W, the zero line in Figure 7 is below the lower ASEE band, which may indicate that the scale drift in W is caused by sources other than random equating error.

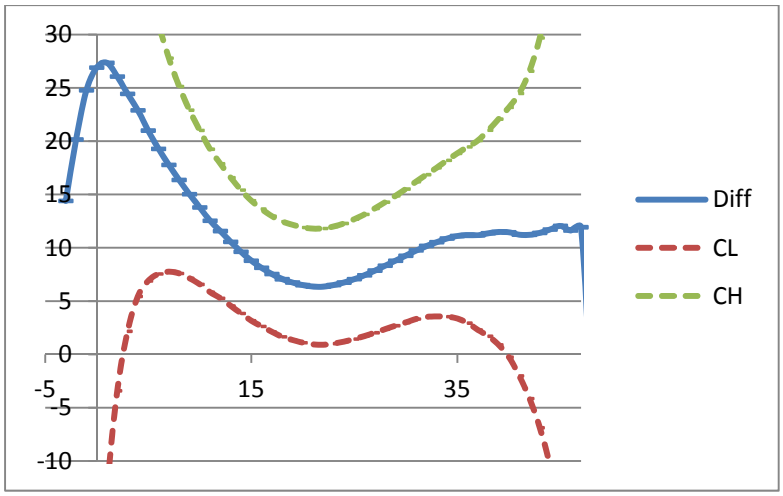

Figure 7. **The difference of the new and original raw-to-scale conversions of OF with the accumulative SEE for W, where CL is the lower boundary and CH is the higher boundary of the SEE band. OF = old form; SEE = standard error of equating; W = Writing.**

Conclusion

This study examines the stability of the SAT Reasoning Test scale from 2005 to 2010. A 2005 OF was administered along with a 2010 NF. A new conversion for OF was derived through direct equipercentile equating. A comparison of the newly derived and the original OF conversions showed that the CR and M scales have experienced moderate upward scale drift (no greater than 5 points each on average) and the drift is within the accumulation of equating errors caused by random sampling. The W scale has experienced a significant upward scale drift (about 11 points on average).

Haberman and Dorans (2011) listed possible sources for scale drift. For example, population shift, test construction shift, violation of equating assumptions, random sampling error, and so on may cause a scale to drift. In this study, we investigated the equating error caused by random sampling and whether the scale drift could be partly explained by the random equating error. Random equating error can accumulate after a series of equatings, which may account for scale drift. It is not clear what has caused the W scale to drift. One possible explanation is that the W measure was first introduced in the SAT Reasoning Test in March 2005, and the OF was originally equated to the March, May, and June 2005 forms; the scale may not have been stabilized at the initial administrations. In contrast, the M and CR scales were

continuations of well-established and stable SAT-Mathematics and SAT-Verbal scales (Haberman, Guo, Liu, & Dorans, 2008).

What shall the testing programs do when a scale drift is detected? Liu, Curley, and Low (2009) recommended a few procedures to investigate the possible shifts/changes in test statistics specifications, content specifications, test-taking populations, administration conditions, statistical procedures, and so on. After that, the testing programs may be able to make adjustments to fix the scale drift problem accordingly. For example, the simultaneous linking/scaling procedure proposed by Haberman (2009) was implemented in the *TOEFL*® program to obtain more accurate and more stable equating results. In extreme cases, adjustment of the scaling of the tests may be necessary to maintain a meaningful scale. Such examples include the SAT recentering (Dorans, 2002), the ACT assessment rescaling (Brennan, 1989), and the ITBS periodical rescaling (Kolen & Brennan, 2004).

In order to help maintain the scale stability and reduce scale drift, researchers have proposed many suggestions, such as constructing parallel test forms, using large equating sample sizes, using a braiding plan (test forms are interweaved to avoid the development of separate strains), and using multiple-linking equating designs (Guo, Liu, Dorans, & Feigenbaum, 2011; Haberman et al., 2008). However, scale drift occurs after a series of equatings even when best practices are followed. It is recommended that testing programs like the SAT monitor the stability of their score scales periodically.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Brennan, R. L. (Ed.). (1989). *Methodology used in scaling the ACT Assessment and P-ACT+*. Iowa City, IA: American College Testing.

Dorans, N. J. (2002). *The recentering of SAT scales and its effect on score distributions and score interpretations* (College Board Report No. 2002-11). New York, NY: College Entrance Examination Board.

Dorans, N. J., Holland, P. W., Thayer, D. T., & Tateneni, K. (2003). Invariance of score linking across gender groups for three *Advanced Placement Program* examinations. In N. J. Dorans (Ed.), *Population invariance of score linking: Theory and applications to Advanced Placement Program examinations* (ETS Research Report No. RR-03-27, pp. 79–118). Princeton, NJ: ETS.

Guo, H. (2010). Accumulative equating error after a chain of linear equatings. *Psychometrika*, *75*, 438–453.

Guo, H., Liu, J., Dorans, N., & Feigenbaum, M. (2011). *Multiple linking in equating and random scale drift* (ETS Research Report No. RR-11-46). Princeton, NJ: ETS.

Haberman, S. (2009). *Linking parameter estimates derived from an item response model through separate calibrations* (ETS Research Report No. RR-09-40). Princeton, NJ: ETS.

Haberman, S. (2010). *Limits on the accuracy of linking* (ETS Research Report No. RR-10-22). Princeton, NJ: ETS.

Haberman, S., & Dorans, N. J. (2011). *Sources of score scale inconsistency* (ETS Research Report No. RR-11-10). Princeton, NJ: ETS.

Haberman, S., Guo, H., Liu, J., & Dorans, N. J. (2008). *Consistency of SAT I: Reasoning Test score conversions* (ETS Research Report No. RR-08-67). Princeton, NJ: ETS.

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate models for discrete test score distributions. *Journal of Educational and Behavioral Statistics*, *25*, 133–183.

Holland, P. W., & Wightman, L. (1982). Section pre-equating. A preliminary investigation. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 271–306). New York, NY: Academic Press.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag.

Liu, J., Curley, E., & Low, A. (2009). *A scale drift study* (ETS Research Report No. RR-09-43). Princeton, NJ: ETS.

Livingston, S. (2004). *Equating test scores (without IRT)*. Princeton, NJ: ETS.

Lord, F. M. (1982). The standard error of equipercentile equating. *Journal of Educational Statistics*, *7*, 156–174.

McHale, F. J., & Ninneman, A. M. (1994). *The stability of the score scale for the Scholastic Aptitude Test from 1973 to 1984* (ETS Statistical Report No. SR-94-27). Princeton, NJ: ETS.

Modu, C. C., & Stern, J. (1975). *The stability of the SAT score scale* (ETS Research Bulletin No. RB-75-9). Princeton, NJ: ETS.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics*, *8*(2), 137–156.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). New York, NY: Macmillan.

Puhan, G. (2009). Detecting and correcting scale drift in test equating: An illustration from a large scale testing program. *Applied Measurement in Education*, *22*, 79–103.

Stewart, E. E. (1966). *The stability of the SAT-Verbal score scale* (ETS Research Bulletin No. RB-66-37). Princeton, NJ: ETS.

Wilks, S. S. (Ed.). (1961). *Scaling and equating College Board tests.* Princeton, NJ: ETS.

Appendix

Figure A1 displays the difference between the new and original raw-to-raw conversions of the studied form OF CR measure (the solid line labeled diff) with the SEE band for the raw difference: difference ±SEE (the two dashed lines).

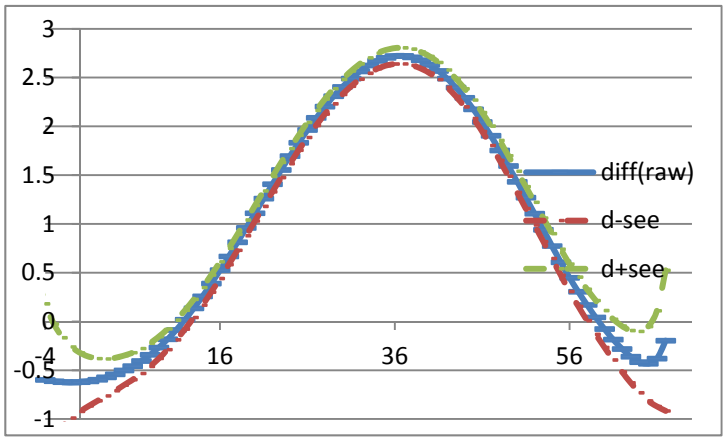

Figure A1. **The difference between the new and original CR raw-to-raw conversions and SEE. CR = Critical Reading; SEE = standard error of equating.**

Figure A2 displays the difference between the new and original raw-to-raw conversions of the OF M measure.

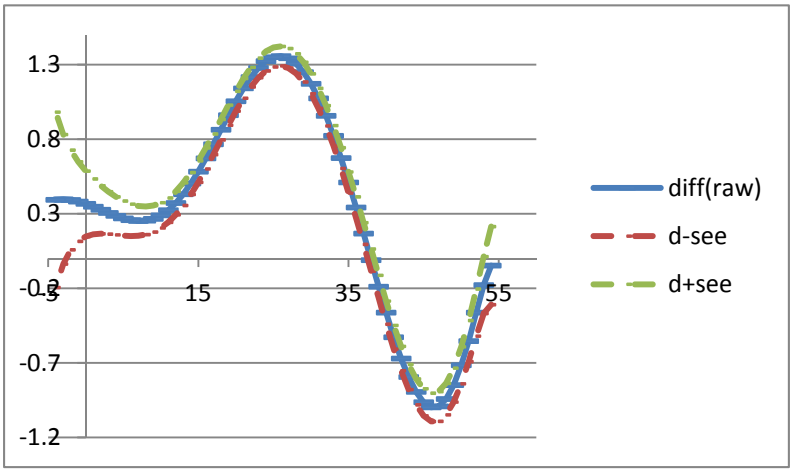

Figure A2. **The difference between the new and original M raw-to-raw conversions and SEE. M = Mathematics; SEE = standard error of equating.**

Figure A3 displays the difference between the new and original raw-to-raw conversions of the OF W measure.

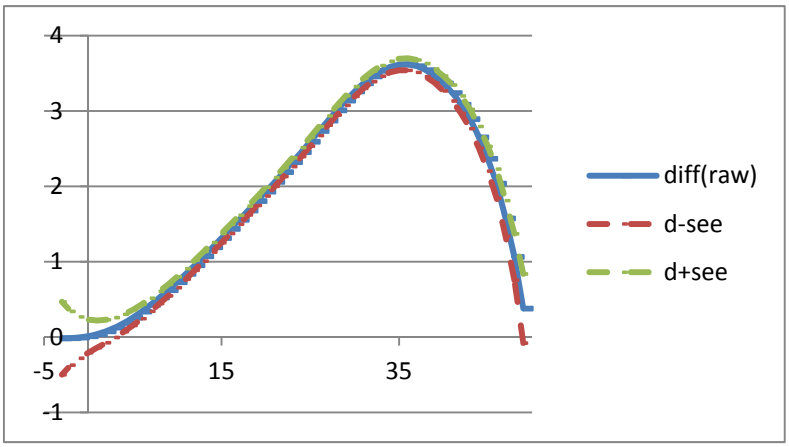

Figure A3. **The difference between the new and original W raw-to-raw conversions and SEE. W = Writing; SEE = standard error of equating.**

In Table A1, the first column is the raw score, the second column is the original conversion, the third column is the newly equated conversion, the fourth column is the unrounded difference between the new and original conversions of the studied form OF, the fifth column is the SEE of raw scores, and the last column is the SEE of scaled scores. In the middle range of the raw scores (9 to 51), the SEE of scaled scores is less than 1, which is negligible considering the SAT scale (200 to 800). The SEE of scaled scores at the very top (Formula Scores 64 to 67) and the very bottom (Formula Scores -3 to 0) are relatively large because of approximation and fewer numbers of examinees at those score levels.

In Table A2, the SEE of the middle range of the raw scores (9 to 46) is less than 1, which is negligible considering the SAT scale (200 to 800). The SEE of scaled scores at the very top (Formula Score 64) and the very bottom (Formula Scores -2 to -1) are relatively large because of approximation and fewer numbers of examinees at those score levels.

In Table A3, the SEE of raw scores in the middle range of the raw scores (7 to 39) is less than 1, which is negligible considering the SAT scale (200 to 800). The SEE of scaled scores at the very top (Formula Scores 48 to 49) and the very bottom (Formula Scores -3 to 0) are relatively large.

Table A1

Comparison of the Old and New Equating Conversions and SEE for Critical Reading

CR FS	Original CR (11/05)	New CR (11/10)	Unrounded difference	SEE (raw)	SEE (scale)
67	868.92	866.53	-2.39	0.72	21.58
66	836.65	833.18	-3.47	0.49	11.95
65	809.58	808.66	-0.92	0.38	8.33
64	787.46	788.18	0.72	0.31	6.11
63	768.50	770.30	1.80	0.27	4.84
62	752.62	754.22	1.60	0.23	3.79
61	738.75	739.90	1.15	0.21	3.06
60	726.32	726.93	0.60	0.19	2.55
59	714.85	715.07	0.22	0.17	2.18
58	704.08	704.08	0.00	0.16	1.92
57	693.88	693.64	-0.24	0.15	1.69
56	684.19	683.78	-0.41	0.14	1.51
55	674.96	674.51	-0.44	0.13	1.36
54	666.11	665.68	-0.42	0.13	1.23
53	657.59	657.23	-0.36	0.12	1.12
52	649.42	649.13	-0.30	0.12	1.04
51	641.56	641.37	-0.19	0.11	0.96
50	633.92	633.83	-0.10	0.11	0.91
49	626.46	626.52	0.05	0.11	0.85
48	619.13	619.38	0.25	0.10	0.80
47	611.91	612.33	0.43	0.10	0.76
46	604.81	605.41	0.61	0.10	0.71
45	597.80	598.59	0.79	0.09	0.68
44	590.88	591.84	0.96	0.09	0.65
43	584.08	585.17	1.09	0.09	0.62
42	577.34	578.57	1.22	0.09	0.60
41	570.69	572.04	1.36	0.09	0.57
40	564.10	565.57	1.47	0.09	0.56
39	557.54	559.10	1.57	0.08	0.54
38	551.03	552.66	1.63	0.08	0.53
37	544.55	546.22	1.68	0.08	0.51
36	538.06	539.79	1.73	0.08	0.50
35	531.59	533.35	1.76	0.08	0.49
34	525.09	526.88	1.79	0.08	0.48
33	518.59	520.40	1.81	0.08	0.47
32	512.06	513.89	1.82	0.08	0.47
31	505.51	507.36	1.86	0.08	0.46
30	498.97	500.84	1.87	0.08	0.46
29	492.38	494.30	1.91	0.08	0.46
28	485.77	487.76	1.98	0.08	0.45
27	479.11	481.16	2.04	0.08	0.46
26	472.36	474.52	2.16	0.08	0.46
25	465.56	467.86	2.30	0.08	0.46
24	458.61	461.09	2.48	0.08	0.47
23	451.57	454.30	2.73	0.08	0.47
22	444.44	447.44	3.01	0.08	0.47
21	437.22	440.53	3.30	0.08	0.49
20	429.88	433.58	3.70	0.08	0.49
19	422.42	426.52	4.11	0.08	0.51
18	414.81	419.41	4.60	0.08	0.53

CR FS	Original CR (11/05)	New CR (11/10)	Unrounded difference	SEE (raw)	SEE (scale)
17	407.00	412.18	5.19	0.08	0.55
16	398.99	404.81	5.82	0.09	0.57
15	390.80	397.31	6.51	0.09	0.60
14	382.37	389.61	7.24	0.09	0.63
13	373.68	381.75	8.07	0.09	0.68
12	364.71	373.67	8.96	0.10	0.73
11	355.31	365.26	9.95	0.10	0.80
10	345.54	356.52	10.98	0.11	0.89
9	335.22	347.34	12.11	0.11	0.99
7	312.68	327.41	14.73	0.13	1.30
6	300.25	316.40	16.15	0.14	1.50
5	286.78	304.54	17.75	0.15	1.81
4	272.12	291.53	19.41	0.17	2.19
3	256.11	277.13	21.01	0.19	2.78
2	238.44	261.04	22.60	0.22	3.56
1	219.61	243.12	23.51	0.25	4.64
0	199.95	223.56	23.61	0.30	5.99
-1	179.89	202.71	22.81	0.36	7.75
-2	159.28	180.37	21.09	0.46	10.52
-3	138.52	157.21	18.69	0.59	13.79

Note. CR = Critical Reading; FS = formula score; new CR = equated Critical Reading conversion; SEE = standard error of equating.

Table A2

Comparison of the Old and New Equating Conversions and SEE for Mathematics

M FS	Original M (11/05)	New M (11/10)	Unrounded difference	SEE (raw)	SEE (scale)
54	798.82	806.23	7.40	0.26	7.03
53	770.18	776.65	6.47	0.19	4.31
52	743.02	751.26	8.24	0.16	2.98
51	724.52	730.97	6.45	0.14	2.06
50	709.99	715.07	5.08	0.12	1.60
49	697.67	701.57	3.90	0.11	1.32
48	686.80	689.64	2.84	0.11	1.15
47	676.88	678.88	2.00	0.10	1.02
46	667.60	668.83	1.23	0.09	0.95
45	658.64	659.29	0.65	0.09	0.90
44	650.00	650.11	0.11	0.09	0.86
43	641.53	641.16	-0.37	0.08	0.84
42	633.14	632.43	-0.72	0.08	0.82
41	624.87	623.86	-1.01	0.08	0.80
40	616.66	615.53	-1.13	0.08	0.77
39	608.45	607.33	-1.11	0.08	0.75
38	600.32	599.20	-1.12	0.07	0.72
37	592.23	591.26	-0.97	0.07	0.69
36	584.19	583.43	-0.76	0.07	0.66
35	576.17	575.68	-0.50	0.07	0.63
34	568.23	568.05	-0.18	0.07	0.60
33	560.33	560.52	0.19	0.07	0.58
32	552.47	553.04	0.57	0.07	0.55

M FS	Original M (11/05)	New M (11/10)	Unrounded difference	SEE (raw)	SEE (scale)
31	544.67	545.57	0.90	0.07	0.53
30	536.90	538.20	1.30	0.07	0.52
29	529.17	530.85	1.68	0.07	0.50
28	521.47	523.51	2.04	0.07	0.49
27	513.78	516.16	2.38	0.07	0.48
26	506.09	508.79	2.70	0.07	0.47
25	498.43	501.41	2.98	0.07	0.47
24	490.74	494.00	3.25	0.07	0.46
23	483.03	486.51	3.47	0.07	0.46
22	475.29	478.97	3.68	0.07	0.47
21	467.52	471.36	3.84	0.07	0.47
20	459.68	463.63	3.95	0.07	0.48
19	451.74	455.81	4.07	0.07	0.50
18	443.75	447.88	4.12	0.07	0.51
17	435.62	439.79	4.17	0.07	0.53
16	427.34	431.56	4.22	0.07	0.56
15	418.85	423.16	4.31	0.07	0.58
14	410.20	414.55	4.35	0.07	0.62
13	401.26	405.70	4.43	0.08	0.66
12	392.06	396.57	4.52	0.08	0.72
11	382.49	387.15	4.65	0.08	0.79
10	372.45	377.31	4.85	0.08	0.86
9	361.91	367.07	5.17	0.09	0.96
8	350.78	356.35	5.57	0.09	1.07
7	339.05	345.12	6.07	0.10	1.20
6	326.67	333.32	6.65	0.11	1.38
5	313.59	320.80	7.21	0.12	1.61
4	299.66	307.41	7.75	0.13	1.89
3	284.75	293.15	8.40	0.14	2.32
2	268.78	277.65	8.87	0.16	2.75
1	251.49	261.18	9.69	0.19	3.48
0	232.72	243.40	10.68	0.22	4.16
-1	214.11	224.86	10.75	0.27	5.34
-2	195.51	205.50	9.98	0.33	6.81
-3	176.38	185.38	9.00	0.43	9.08

Note. M = Mathematics; FS = formula score; new M = equated Mathematics conversion; SEE = standard error of equating.

Table A3

Comparison of the Old and New Equating Conversions and SEE for Writing

W FS	Original W (11/05)	New W (11/10)	Unrounded difference	SEE (raw)	SEE (scale)
49	844.26	800.00	-44.26	0.46	19.62
48	806.50	800.00	-6.50	0.31	8.44
47	778.23	790.16	11.94	0.24	4.96
46	755.50	767.13	11.63	0.19	3.40
45	736.16	748.22	12.06	0.16	2.56
44	719.13	730.85	11.72	0.14	2.02
43	703.92	715.33	11.41	0.13	1.66
42	689.91	701.13	11.22	0.12	1.39

W FS	Original W (11/05)	New W (11/10)	Unrounded difference	SEE (raw)	SEE (scale)
41	676.54	687.77	11.23	0.11	1.20
40	663.63	675.09	11.46	0.10	1.06
39	651.51	662.99	11.49	0.09	0.96
38	640.07	651.44	11.37	0.09	0.88
37	629.13	640.31	11.18	0.08	0.81
36	618.27	629.47	11.20	0.08	0.75
35	607.69	618.80	11.11	0.08	0.70
34	597.36	608.23	10.87	0.07	0.67
33	587.24	597.78	10.53	0.07	0.63
32	577.26	587.45	10.19	0.07	0.61
31	567.45	577.22	9.76	0.07	0.59
30	557.81	567.07	9.26	0.07	0.57
29	548.20	556.99	8.79	0.06	0.55
28	538.66	546.98	8.33	0.06	0.54
27	529.16	537.02	7.87	0.06	0.53
26	519.66	527.08	7.43	0.06	0.52
25	510.12	517.18	7.06	0.06	0.51
24	500.59	507.33	6.74	0.06	0.50
23	491.07	497.58	6.51	0.06	0.50
22	481.52	487.88	6.36	0.06	0.50
21	471.87	478.24	6.37	0.06	0.49
20	462.14	468.64	6.50	0.06	0.49
19	452.32	459.07	6.76	0.06	0.49
18	442.50	449.56	7.06	0.06	0.49
17	432.57	440.11	7.54	0.06	0.49
16	422.59	430.72	8.13	0.06	0.50
15	412.61	421.40	8.79	0.06	0.51
14	402.48	412.11	9.63	0.06	0.53
13	392.23	402.80	10.57	0.06	0.55
12	381.87	393.45	11.58	0.07	0.57
11	371.45	383.99	12.54	0.07	0.61
10	360.60	374.40	13.80	0.07	0.66
9	349.57	364.60	15.03	0.08	0.72
8	338.15	354.51	16.37	0.08	0.80
7	326.22	343.99	17.78	0.09	0.91
6	313.59	332.88	19.29	0.09	1.07
5	299.97	320.97	21.00	0.10	1.28
4	285.15	308.05	22.90	0.12	1.59
3	269.36	293.80	24.44	0.13	2.02
2	251.81	277.88	26.07	0.15	2.68
1	232.43	259.79	27.36	0.18	3.66
0	212.09	239.00	26.91	0.22	5.06
-1	190.71	215.50	24.78	0.27	7.10
-2	169.21	189.38	20.17	0.36	10.10
-3	146.66	161.06	14.40	0.48	13.71

Note. W = Writing; FS = formula score; new W = equated Writing conversion; SEE = standard error of equating.