An Investigation of ACT Equating Stability

Dongmei Li, PhD

Equating is the statistical procedure that adjusts scores from a given ACT® test form to make them interchangeable with scores from other ACT test forms, which may differ slightly in difficulty. Stability is an important property of equating results because it ensures that score meaning is consistent over time and independent of the examinee samples used for equating. ACT regularly conducts "stability check" analyses by re-equating forms that were equated previously. This report describes one such analysis conducted after the February 2020 ACT administration. Overall, results indicated that differences in equating results were within the range of what would be expected due to random sampling error. Therefore, this study provides evidence supporting the stability of ACT equating results.

Background

Equating for the ACT test uses a random groups design. That is, new ACT forms and an anchor form are spiraled to examinees such that the samples taking the different forms are randomly equivalent in ability and other characteristics. With this design, observed differences in test scores between the groups taking different forms can be attributed to form differences instead of ability differences. The anchor form, which was equated previously, helps ensure continuity of the score scale over time and accommodates changes in the distribution of ability between equating studies.

For any equating study, results can potentially be impacted by two general sources of error: random error and systematic error (Kolen & Brennan, 2014). Random error, which results from using samples rather than the whole population, can be estimated with analytical or re-sampling methods, and it can be reduced by increasing sample size. Systematic error may be introduced when the data collection design is not implemented properly or when assumptions of the equating methods are violated. Unlike random error, systematic error, which is more difficult to detect and quantify, can persist even when sample sizes are increased.

Equating results are impacted by both random and systematic errors, and large-scale testing programs should monitor the stability of equating results to ensure that score interpretation is consistent over time. One approach is re-equating a form that was previously equated. The overall equating error can be evaluated by comparing equating results from the different equating studies, though the two sources of error cannot be clearly disentangled.

ACT periodically conducts stability check studies through re-equating of forms. In these studies, a form equated in earlier years is re-equated. This report describes results of the stability check study conducted in February 2020, in which a form equated in October 2018 was re-equated.

Two forms equated in October 2018 were included again in February 2020 equating—one form (Form A) was included as an equating anchor, and the other (Form S) was included for the purpose of conducting a stability check. The stability check form was used to evaluate ACT equating in two ways:

1. The impact of different equating samples was investigated by comparing equating results for the stability check form (Form S) between October 2018 and February 2020.

2. The impact of using a different equating anchor form was investigated by comparing equating results from the use of the two different anchor forms: the actual anchor form (Form A) and the stability check form (Form S).

Provided below is a brief description of procedures and findings from these two investigations.

Part 1: Equating Results Comparison for the Stability Check Form

The use of anchor forms across equating studies accommodates differences in equating samples, so there is no requirement that the equating samples be equivalent across years. Despite this, ACT takes care to ensure the equating samples across years are highly similar (in terms of academic achievement on the ACT) to minimize the possibility that equating results are sensitive to the distribution of ability in the equating sample. One reason ACT conducted a stability check study in February 2020 was that this was the first time that equating was conducted using data from the February national test administration.

Table 1 shows the means and standard deviations of the subject test scale scores (on the 1–36 ACT scale) and raw scores (number correct) for the two forms and the differences between the forms across years (in the “Form S-A” columns). It also shows the sample differences for each form between the two years (in the “2020-2018” rows). The scale scores for the stability check form (Form S) were based on the operational conversion tables obtained from the 2018 equating study. T-tests were conducted for the difference values in the “Form S-A” columns and the “2020-2018” rows.

The equating sample in February 2020 was slightly higher in ability than the sample in October 2018, though one form suggested a slightly greater difference between the two years than the other form. This is apparent when comparing the raw score and scale score means of the two forms between the two years and by the positive 2020-2018 differences between the means for each form. For example, based on the scale score statistics of the English test in Form S, the 2018 sample had a mean of 21.09, and the 2020 sample had a mean of 21.45—a statistically significant difference of 0.36 ($p < .05$). The other three subject tests on this form also showed statistically

significant increases from the 2018 to the 2020 equating samples. Form A also showed higher subject test means for the 2020 sample than the 2018 sample, yet none of the differences were statistically significant. The relatively higher performance of the 2020 equating sample was not expected to have any systematic impact on equating results due to the use of anchor forms across years.

Based on comparisons of the raw score means between the two forms in each year, the math test on Form S was significantly harder than the math test on Form A in both the 2018 and the 2020 equating samples ($p < .001$), but none of the other subject tests exhibited statistically significant raw score differences between the two forms. This consistency of relative form differences across years provides a necessary precondition for the stability of equating results.

Table 1 also shows some sampling differences between the two years that may have impacted the equating results of Form S. Note that, if there was no sampling error nor equating error, one should expect the scale score mean differences to be close to zero and the raw score differences between forms to be similar across years for each subject test. There was consistency in that only the average math test scores were significantly different between the two forms, but the observed differences between the two forms differed slightly across years. For example, the Form S raw score mean for English was higher than Form A by 0.32 in 2018 but by 0.64 in 2020. In addition, the scale score means for reading had a statistically significant difference ($p = .049$) between the two forms in 2020. These differences were manifestations of random and other errors that may have contributed to equating differences.

Table 1. Sample and Form Comparisons Across Years

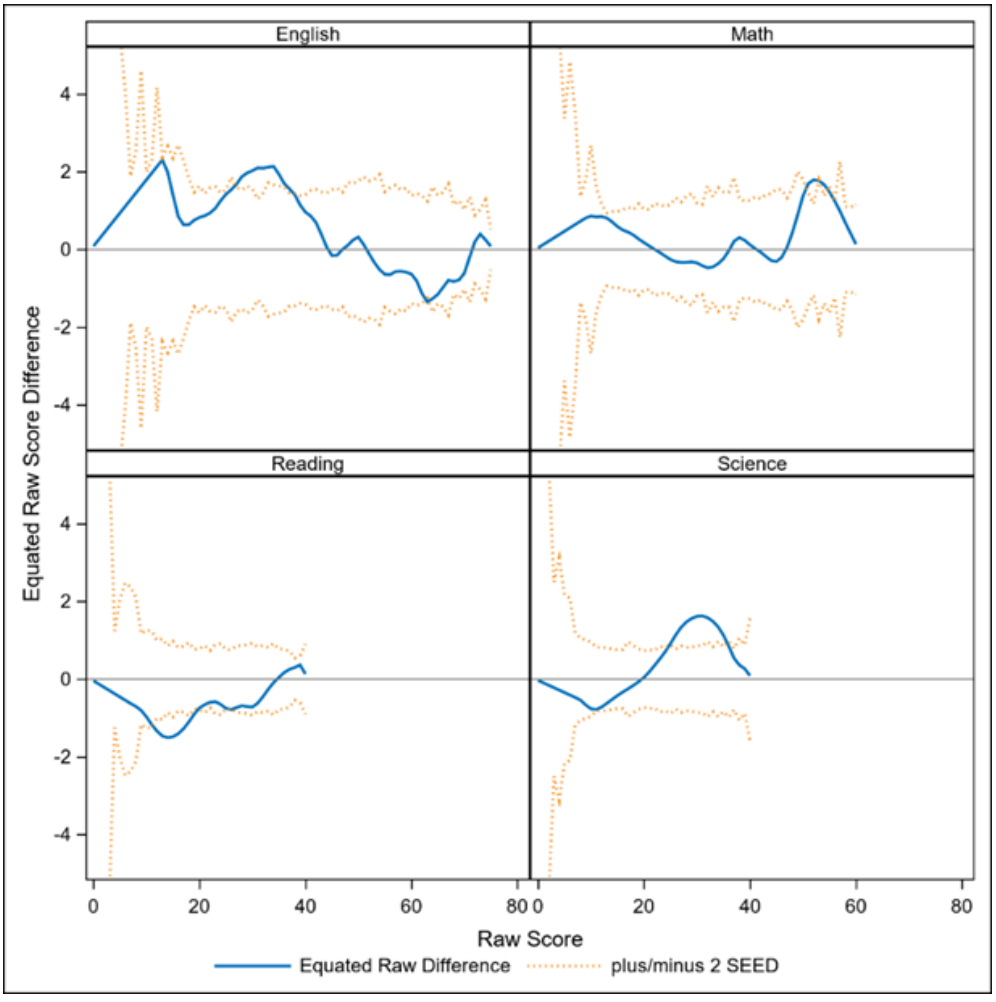
		Scale Score			Raw Score		
		Form S	Form A	Form S-A	Form S	Form A	Form S-A
2018	English	21.09	21.09	0.00	46.96	46.63	0.32
	Math	21.20	21.19	0.01	30.48	31.84	-1.36***
	Reading	22.34	22.28	0.06	24.87	24.88	-0.01
	Science	21.61	21.57	0.04	22.70	22.87	-0.17
2020	English	21.45	21.28	0.17	47.59	46.95	0.64
	Math	21.54	21.42	0.12	31.05	32.17	-1.12***
	Reading	22.78	22.42	0.36*	25.29	24.99	0.30
	Science	21.99	21.83	0.16	23.11	23.18	-0.07
2020-2018	English	0.36*	0.19	0.17	0.63	0.32	0.32
	Math	0.34*	0.23	0.11	0.57	0.33	0.24
	Reading	0.44*	0.14	0.30	0.42*	0.11	0.31
	Science	0.38**	0.26	0.12	0.41*	0.31	0.10

Note: * $p < .05$, ** $p < .01$, *** $p < .001$

Equating results for the stability check form were compared between the two equating events (October 2018 and February 2020) in terms of conversion table differences and group means. In the equating process, raw scores on a new form are first equated to the raw scores on the anchor form, then the anchor form raw to unrounded scale score conversions are used to derive the unrounded scale scores for the new form. Finally, the scale scores are rounded to integers for the purpose of score reporting. For this investigation, plots were generated to illustrate the differences in equated raw scores, unrounded scale scores, and rounded scale scores at each raw score point between the two equating events. Calculations included the mean difference in reported scale scores when applying the two different conversions for the group taking the test form in February 2020.

Figure 1 presents the equated raw score differences between 2020 and 2018 along with error bands (the dotted lines) representing plus or minus two standard errors of equating differences (SEED) for the four subject tests. Assuming that equating errors were not correlated between different equatings, the SEED at each raw score point was calculated as the square root of the sum of the squared standard error of equating (i.e., the equating error variance) at that score point across the two equatings. With few exceptions, the equated raw score differences between the two equatings for each subject test were within two standard errors. These results indicate that the differences between equating results from 2018 and 2020 were mostly within the range of differences that would be expected due to random error.

Figure 1. Differences in Equated Raw Scores



There are 76, 61, 41, and 41 raw score points for English, math, reading, and science, respectively, but the scale scores for all four subject tests range from 1 to 36. Since there are more raw score points than scale score points for the ACT test, differences in the equated raw scores do not always impact scale scores. Figure 2 presents the unrounded scale score differences between the two equatings at each raw score point (i.e., differences between raw-to-unrounded scale score conversions). With few exceptions, the absolute differences for the unrounded scale scores were below 0.5 scale score points, and the largest rounded scale score differences were one scale score point. Note that 0.5 often serves as a point of reference because differences greater than or equal to 0.5 in magnitude will result in different reported scale scores on the 1–36 scale (i.e., the "difference that matters" criterion). The percentage of raw score points with unrounded scale score differences smaller than 0.5 were 83%, 100%, 83%, and 95% for English, math, reading, and science, respectively. Larger differences generally occurred near the low and high ends of the score scale, where random errors are greater due to smaller sample sizes.

Figure 2. Unrounded Scale Score Differences

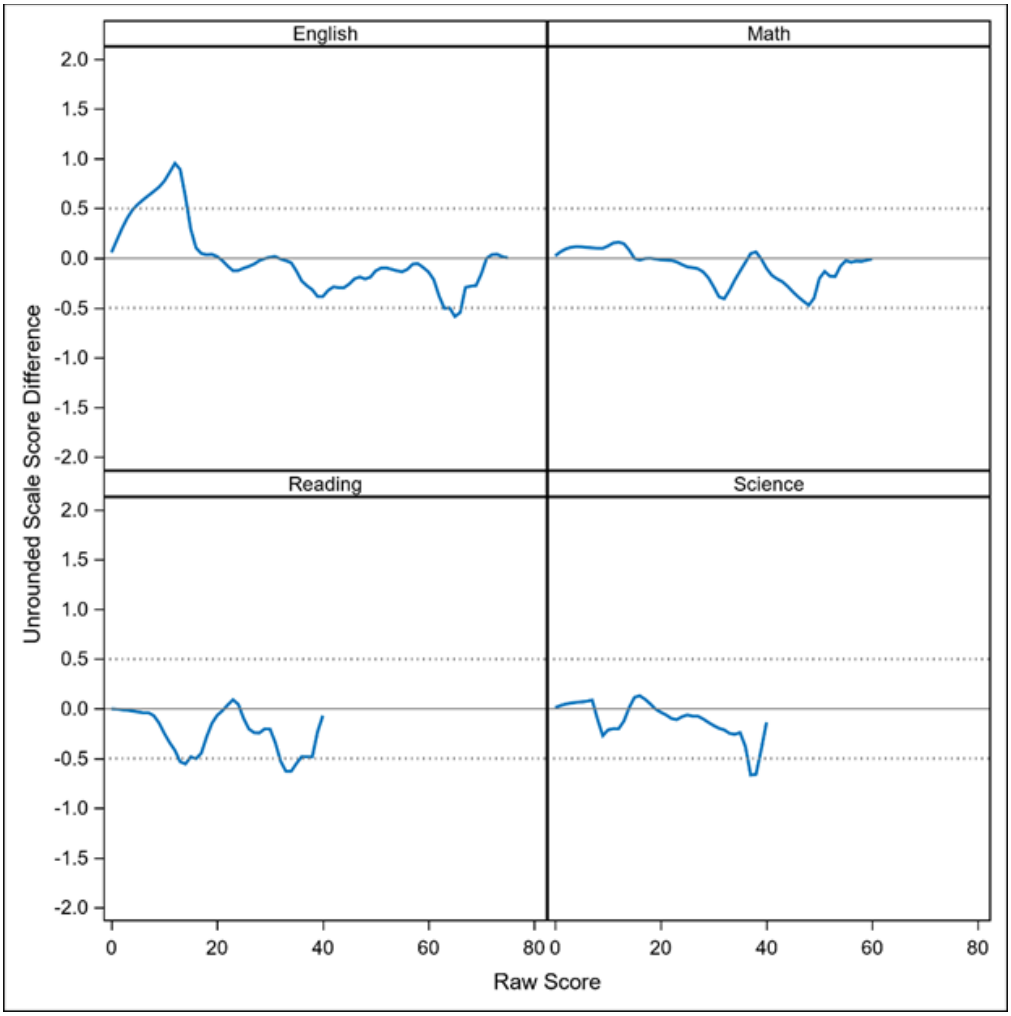
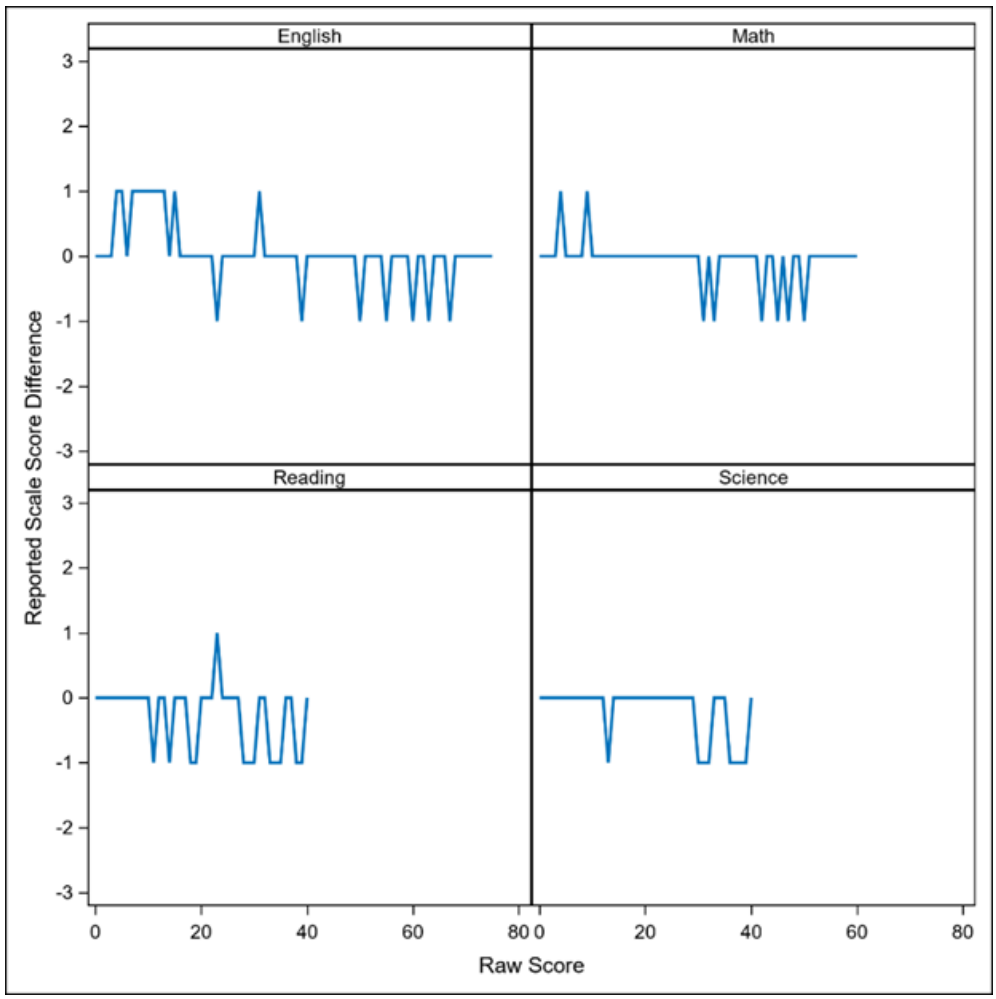


Figure 3 presents the rounded scale score differences between the two equatings at each raw score point. The maximum difference was plus or minus 1 score point in each of the four subject tests. Since it is the rounded scale scores that are used for score reporting, Figure 3 indicates the potential impact on individual scores at different score levels. Whereas the scale scores may change by one point if conversions obtained from a different year were used, the majority of the raw-to-scale score conversions stayed the same between the two equating studies. The percentage of raw score points that had the same scale score conversions were 76%, 87%, 68%, and 80% for English, math, reading, and science, respectively.

Figure 3. Rounded Scale Score Differences



To examine how different equating results impacted group means, the raw-to-scale score conversion tables of the stability check form from 2018 and 2020 were applied to the group taking this form in February 2020. This approach holds the sample constant to evaluate the impact of applying different conversion tables. Table 2 presents the means and standard deviations of the group, as well as the differences between these statistics, when applying the raw-to-scale score conversion tables from 2018 and 2020. As shown in Table 2, the group means changed slightly on all subject tests when applying the different conversion tables, which resulted in a decrease of .19 in the ACT Composite score mean (average of English, math, reading, and science) when applying the conversions obtained in 2020 compared to 2018.

Table 2. Descriptive Statistics and Differences

	2018 Conversions		2020 Conversions		Difference	
	Mean	SD	Mean	SD	Mean	SD
English	21.45	6.55	21.34	6.47	-0.11	-0.08
Math	21.54	5.72	21.42	5.62	-0.12	-0.10
Reading	22.78	6.71	22.45	6.58	-0.33	-0.13
Science	21.99	5.57	21.81	5.39	-0.18	-0.17
Composite	22.06	5.62	21.88	5.50	-0.19	-0.12

Part 2: Equating Results Comparison Using Different Anchor Forms

In February 2020, 16 new forms were equated. To investigate the impact of using different equating anchor forms, the 16 new forms were equated twice—first using the designated anchor form (Form A) and then using the stability check form (Form S) as the anchor. ACT uses equipercentile equating with post-smoothing for the ACT test equating. In operational equating, smoothing values are chosen carefully by multiple psychometricians. For this investigation, a smoothing value of .05 was used for all forms.

Note that the equating differences observed in Part 1 of this investigation reflected the use of different anchors and different equating samples. In Part 2, observed differences in equating results reflected only the use of different anchor forms.

Differences between conversions using different anchor forms were plotted and examined for the equated raw scores, the unrounded scale scores, and the rounded scale scores of each subject test on each test form. Figures 4 through 6 present these plots for two randomly selected new forms to illustrate the main findings when examining the plots for all the new forms. Observations made from these two forms are also true for the other new forms.

Figure 4 shows the equated raw score differences between equatings using the two different anchors for two new forms (Form B and Form C). The following observations can be made from Figure 4. First, the equated raw score differences were mostly within two SEED for all subject tests in each form. Second, the patterns of these difference plots for each subject test were very similar across forms.

Figure 4. Differences in Equated Raw Scores for two Forms when Equated Using Different Anchors

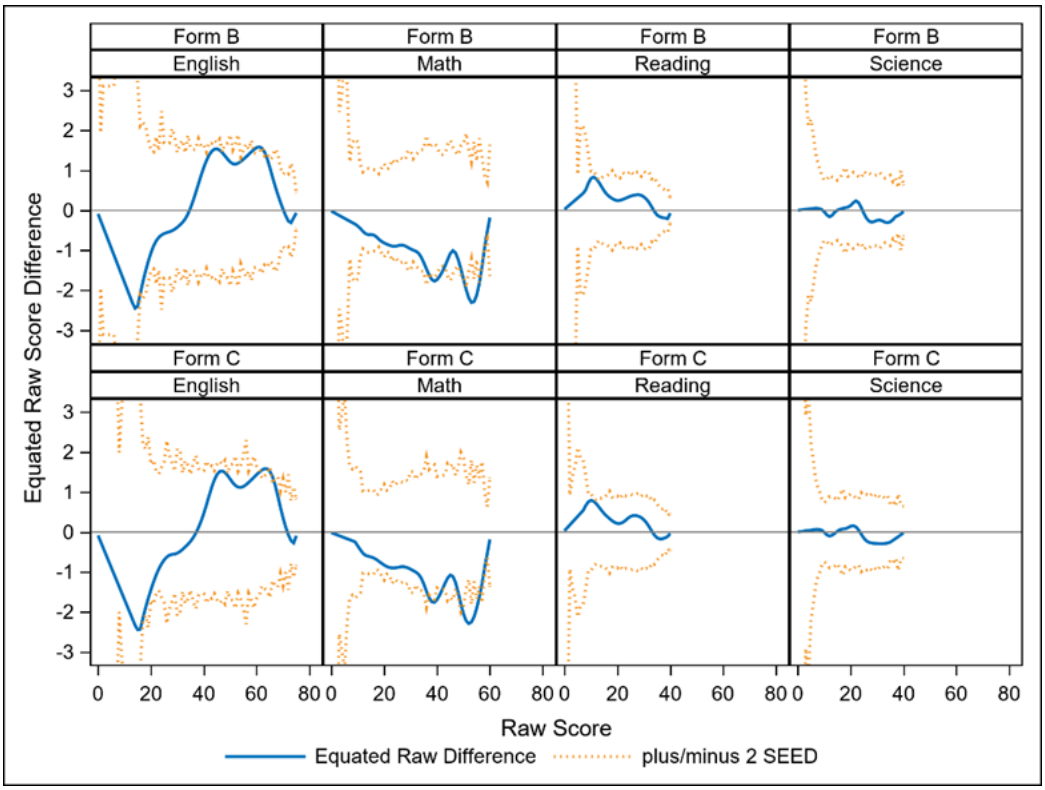
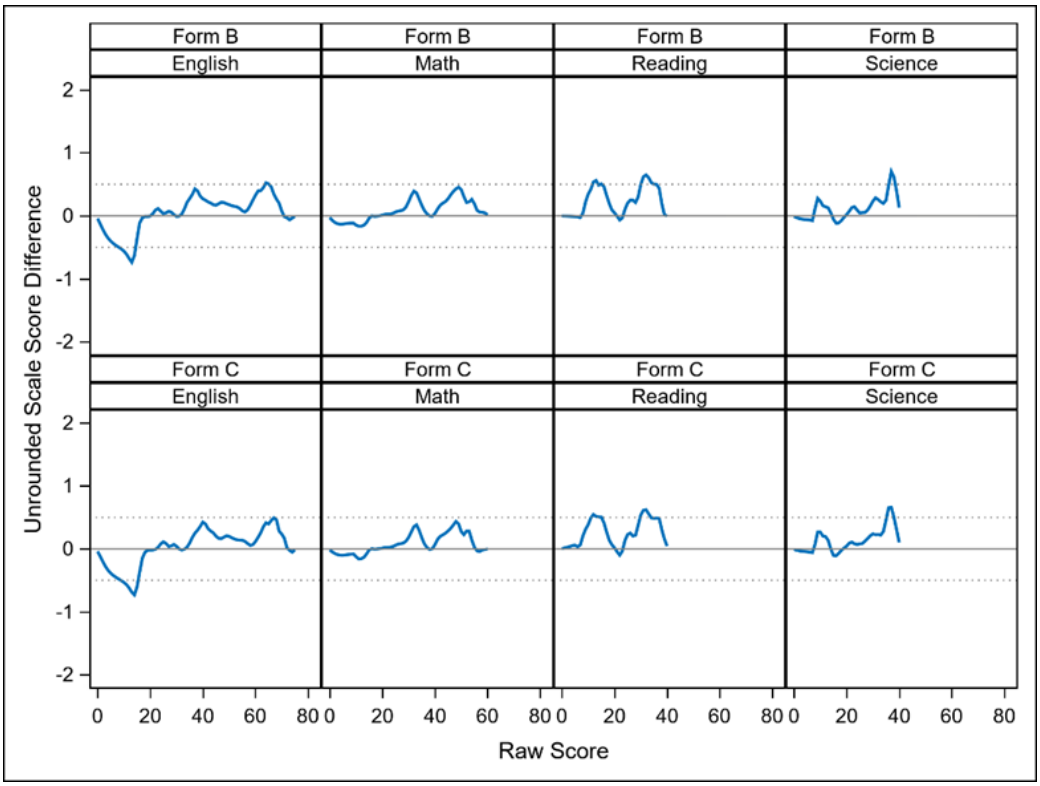


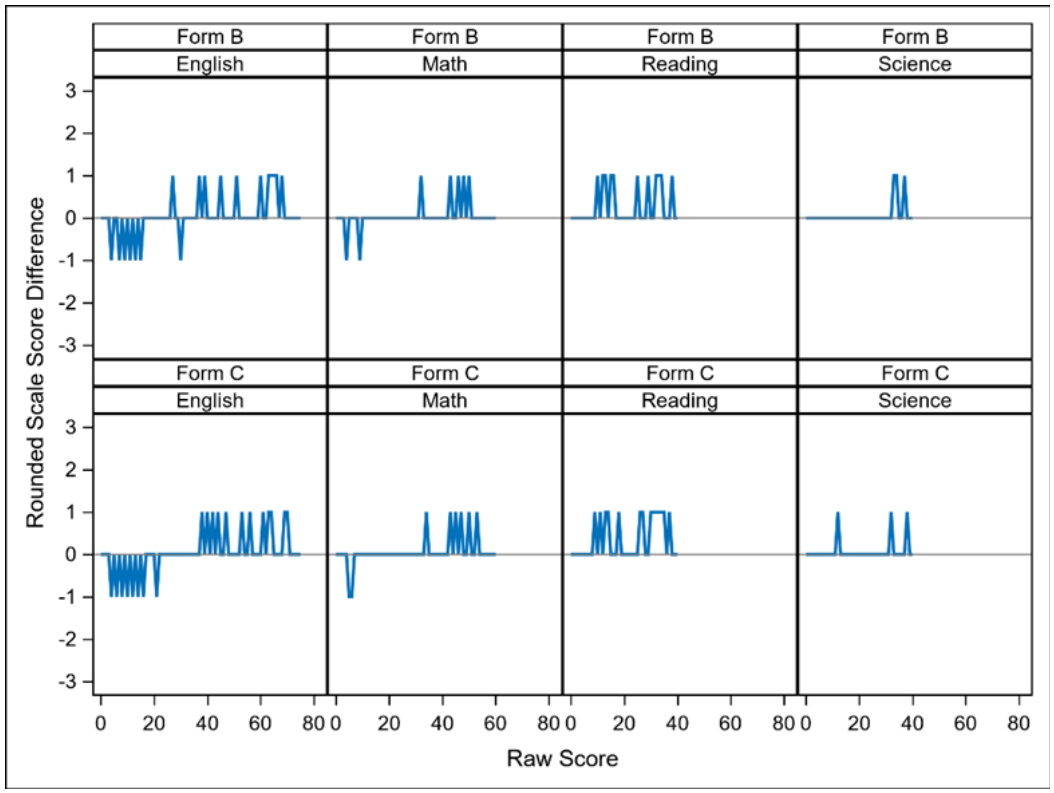
Figure 5 shows the unrounded scale score differences between equatings using the two anchor forms for the two new forms. The magnitudes of differences were mostly below 0.5, which is often considered a criterion for differences that matters. Again, the patterns of these difference plots for each subject were very similar across forms.

Figure 5. Differences in Unrounded Scale Scores for two Forms when Equated Using Different Anchors



The rounded scale score differences for the two forms are presented in Figure 6. Though the unrounded scale score differences in Figure 5 for each subject test were very similar across the two forms, the rounded differences in Figure 6 showed slightly more differences between the two forms because of rounding. The maximum rounded scale score differences across all forms was no more than plus or minus one score point.

Figure 6. Differences in Rounded Scale Scores for two Forms When Equated Using Different Anchors



To summarize the differences in the **unrounded** scale score conversions across all 16 new forms, the percentages of raw score points with unrounded scale score absolute differences below 0.5 were calculated for each subject test of each form. The ranges of percentages across forms were 86-95% for English, 100-100% for math, 78-85% for reading, and 93-95% for science.

To summarize the differences in the **rounded** conversions across all 16 new forms, the percentages of raw score points with no changes in the rounded scale scores were calculated for each subject test of each form. The ranges of percentages across forms were 72-83% for English, 80-92% for math, 63-83% for reading, and 76-95% for science.

Table 3 presents the differences in group means for students taking each of the test forms when conversion tables obtained using different anchor forms were applied. On average, the group mean differences across all forms were similar to the group mean differences observed for the stability check form shown in Table 2.

Table 3. Differences in Group Means

Form	English	Math	Reading	Science	Average
B	0.20	0.10	0.30	0.06	0.16
C	0.23	0.11	0.41	0.06	0.20
D	0.13	0.16	0.43	0.15	0.22
E	0.22	0.11	0.19	-0.03	0.12
F	0.17	0.09	0.15	0.09	0.12
G	0.20	0.16	0.39	0.13	0.22
H	0.07	0.13	0.18	0.12	0.12
I	0.09	0.09	0.28	0.19	0.16
J	0.14	0.07	0.38	0.07	0.16
K	0.14	0.11	0.35	0.10	0.17
L	0.18	0.19	0.34	0.04	0.19
M	0.12	0.13	0.24	0.06	0.14
N	0.11	0.15	0.32	0.24	0.21
O	0.17	0.17	0.33	0.01	0.17
P	0.13	0.10	0.31	0.07	0.15
Q	0.10	0.07	0.44	0.17	0.19
Average	0.15	0.12	0.31	0.10	0.17

Conclusions and Discussion

Equating is a statistical process used to adjust scores for test forms that differ slightly in difficulty so that the test forms can be used interchangeably. Since equating is usually conducted using samples instead of the whole test population, equating error is expected. Random errors can be estimated using analytical or re-sampling methods, but stability check studies through re-equating previously equated forms allow the detection of equating errors that may exist beyond random errors. For that reason, large-scale testing programs usually conduct such stability checks periodically.

This report summarized major analyses and results from the February 2020 ACT stability check study in which a form equated in October 2018 was re-equated. Besides random sampling error, some other differences between the 2020 and 2018 equatings of that form might also have contributed to the equating result differences. For example, the 2020 equating results for Form S involved a longer equating than that the 2018 results. That is, in 2020, Form S was equated to the 2018 anchor form (say Form Y) through Form A, but it was directly equated to Form Y in 2018. In addition, the 2020 equating samples were from the February national test administration while the 2018 equating samples were from the October national test administration, and there may have been unknown differences between examinees in these different test administrations.

Examination of comparisons of conversions in terms of equated raw scores, unrounded scale scores, and rounded scale scores, and comparisons of group means when the different conversions were used resulted in several useful observations. First,

the equating differences were mostly within two SEED, indicating that the observed differences between the equating results were mostly in the range of what would be expected from random equating error due to sample differences. Second, the maximum impact of the different equating results on individual scores was no more than one score point for each subject test, and its impact on group means was less than 0.2 on the Composite score scale.

The study also compared equating results using different anchors for 16 forms. The major finding was that using one anchor versus another seemed to have a similar impact on the equating results for all new forms in terms of equated raw scores and unrounded scale scores, yet the final raw-to-scale score conversions of each form may be impacted to different extents because of rounding. Similar to findings when comparing the equating results for the stability check form across two years, the equated raw score differences between results using different anchors were mostly within two SEED, the rounded scale differences were no more than one score point, and the group mean differences were no more than 0.2 on the Composite score scale.

The February 2020 stability check is one example of similar studies that ACT periodically conducts to monitor the stability of scores. The small differences between equating results using different samples or different anchor forms found in this study were consistent with results from ACT's previous internal routine stability checks and major findings from earlier studies investigating ACT equating using different samples (e.g., Harris & Kolen, 1986). These studies provide empirical evidence for the stability of ACT scores and the population invariance property of test equating in general (e.g., Angoff & Cowell, 1986; Dorans & Holland, 2000; Kolen, 2004). On the other hand, the small differences observed in this study and larger differences observed in other studies (e.g., Guo, Liu, Curley, & Dorans, 2012) suggest the need for caution against over-interpretation of small differences at the group level.

References

Angoff, W. H., & Cowell, W. R. (1986). An examination of the assumption that the equating of parallel forms is population-independent. *Journal of Educational Measurement, 23*(4), 327–345.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37*(4), 281–306.

Guo, H., Liu, J., Curley, E., & Dorans, N. (2012). *The stability of the score scales for the SAT Reasoning Test™* from 2005 to 2010. Princeton, NJ: Educational Test Service.

Harris, J. D., & Kolen, J. M. (1986). Effect of examinee group on equating relationships. *Applied Psychological Measurement, 10*(1), 35–43.

Kolen, M. J. (2004). Population invariance in equating and linking: Concept and history. *Journal of Educational Measurement, 41*(1), 3–14.

Kolen, M. J., & Brennan, R. L. (2014). Test equating, scaling, and linking: *Methods and practices* (3rd Edition). New York, NY: Springer.

Dongmei Li, PhD

Dongmei Li is a lead psychometrician in Assessment Transformation at ACT specializing in test equating, scaling, and growth modeling.