Research Report

Assessing Mode Effects of At-Home Testing Without a Randomized Trial

ETS RR–21-10

Sooyeon Kim
Michael Walker

December 2021

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. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

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.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Assessing Mode Effects of At-Home Testing Without a Randomized Trial

Sooyeon Kim & Michael Walker

ETS, Princeton, NJ

In this investigation, we used real data to assess potential differential effects associated with taking a test in a test center (TC) versus testing at home using remote proctoring (RP). We used a pseudo-equivalent groups (PEG) approach to examine group equivalence at the item level and the total score level. If our assumption holds that the PEG approach removes between-group ability differences (as measured by the test) reasonably well, then a plausible explanation for any systematic differences in performance between TC and RP groups that remain after applying the PEG approach would be the operation of test mode effects. At the item level, we compared item difficulties estimated using the PEG approach (i.e., adjusting only for ability differences between groups) to those estimated via delta equating (i.e., adjusting for any systematic differences between groups). All tests used in this investigation showed small, nonsystematic differences, providing evidence of trivial effects associated with at-home testing. At the total score level, we linked the RP group scores to the TC group scores after adjusting for group differences using demographic covariates. We then compared the resulting RP group conversion to the original TC group conversion (the criterion in this study). The magnitude of differences between the RP conversion and the TC conversion was small, leading to the same pass/fail decision for most RP examinees. The present analyses seem to suggest little to no mode effects for the tests used in this investigation.

Keywords At-home testing; remote proctoring; test mode; pseudo-equivalent group; delta equating

doi:10.1002/ets2.12323

Many testing programs introduced a solution that allowed examinees to take the test at home or in another secure location while test centers were closed due to the coronavirus (COVID-19) pandemic. At-home testing offers examinees the same tests offered in test centers, but the testing session is proctored remotely. Availability and convenience are the main benefits of remotely proctored (RP) testing for examinees. Online proctored test delivery makes it possible to take tests at home, thereby saving examinees significant time and expense, particularly when examinees must travel to a test center (TC) that could be hundreds of miles away or even in another country.

RP testing may also have some negatives from the examinee's perspective. For instance, RP requires that the examinee remain in view of the computer camera, thereby restricting the examinee's movement and ability to take breaks. Potential examinees may lack basic remote testing requirements such as a quiet testing environment, sufficient computer hardware and software, adequate internet connectivity, and needed testing accommodations. Some examinees may view RP as an invasion of privacy and may prefer a TC to at-home testing.

From the perspective of the testing organization, the introduction of RP also raises some concerns. *The Standards for Educational and Psychological Testing* lists testing context as a major threat to fair and valid interpretations of test scores (American Educational Research Association et al., 2014). Such threats may include test mode effects and test security concerns. The situation is made more complex by examinee self-selection into TC versus RP testing. It may be that examinees testing under one mode outperform examinees testing under the other purely because of differences in ability. Testing programs have the responsibility to ensure that the results of testing under both modalities are comparable[1] by separating sources of construct-irrelevant variance (e.g., environmental and hardware issues) from construct-relevant variance (e.g., knowledge, skills, and abilities). In the current case, such assurance requires empirical assessment of potential mode effects of at-home testing proctored remotely.

The best approach to investigating test mode effects would be to use a random-groups design to equate a test form given under two different administration modes. A large sample of examinees would take the test under each mode. A counterbalanced single group design could also be used, requiring a much smaller sample of examinees (Eignor, 2007).

Corresponding author: S. Kim, E-mail: skim@ets.org

In either case, the examinees would be randomly selected from the actual test-taking population. The goal of those data collection designs is to eliminate any group difference in ability in order to assess differences in performance caused by test mode. However, collecting the data for such an equating would be not only expensive but also time-consuming. Another problem is that participants may have strong preferences about the mode in which they test. The only way to ensure a truly random sample may be to give the tests in an unofficial administration (i.e., where scores do not count), in which case examinees might be less motivated and might behave differently with respect to the test than they would in an operational administration. Thus, a randomized trial is not a workable option for many testing programs.

Statistical approaches have been proposed to adjust for group differences in ability on the measured construct in situations where no common items are available across the two forms to be equated. Puhan et al. (2007) used propensity scores derived from all available background variables to match the examinees taking a computer-based test to those taking a paper-and-pencil test to assess the comparability of paper-and-pencil and computer-based versions of a certification test. The authors noted that the effectiveness of propensity matching depends on the nature of the covariates included in the procedure.

In another study, Longford (2015) linked 2 years' tertiary school admission tests by forming pairs of examinees across the two cohorts. Because no anchor was available and the two cohorts of examinees were nonequivalent, Longford matched the groups on propensity scores derived from a set of background variables. As did Puhan et al. (2007), the author emphasized that the success of the propensity score approach would require a sufficiently rich set of background variables.

In a third study, Wiberg and Bränberg (2015) used real data from a college admission test to compare the scores derived from three different linking designs: (a) a nonequivalent groups with covariates design, (b) an equivalent groups design with neither covariates nor anchor items, and (c) a nonequivalent groups with anchor test design. The authors concluded that covariates would be useful in practice because not all standardized tests have anchor tests (e.g., because of test security concerns).

Livingston (2014) tried two procedures that used demographic information to achieve groups equivalent in ability, methods he referred to as "demographically adjusted groups" (DAG). In one procedure, Livingston based examinee weights on the linear composite of demographic variables that was most highly correlated with test scores. In the other procedure, he used propensity scores to determine the weights. He found that although both procedures greatly reduced the ability difference between the groups, neither procedure was adequate for creating equivalent groups.

Haberman (2015) proposed pseudo-equivalent groups (PEG) linking, which weights the individual examinees to produce two groups that are statistically matched on demographics. To the extent that the demographic variables are related to group ability, the PEG approach adjusts for group ability differences. PEG and DAG are similar in their nature. The PEG procedure uses examinee background information to construct sample weights via minimum discriminant information adjustment (MDIA). The weights are determined by minimizing a discriminant information function formed from the examinees' background variables (e.g., age, gender, education, major of highest education, job type, a purpose for taking the test, number of times the test was taken previously). The sample weights transform nonequivalent groups of examinees into groups that resemble equivalent groups on selected background variables and thereby potentially reduce group differences in ability.

Several psychometricians at ETS conducted research to assess the practical implications of PEG for large-scale assessment programs using either real or simulated data (see Kim & Lu, 2018; Lu & Guo, 2018; Oh et al., 2015; Xi et al., 2015). A recent empirical investigation (Kim, 2020) using certification tests demonstrated how the PEG approach could be used to explore potential mode effects across RP and TC delivery. As with other studies using background variables, this study showed that the success of the weighting to remove differences in ability depends on the available background variables and their relationship to ability.[2]

The idea of using variables other than scores on an anchor test as a method to adjust for group ability difference is not new (see, e.g., Livingston et al., 1990; Mislevy et al., 1993; Wright & Dorans, 1993). In the absence of anchor tests, then, collateral information has been used for test score linking either as surrogate anchors (Wallin & Wiberg, 2019) or as weighted samples to approximate equivalent groups (Haberman, 2015; Livingston, 2014; Longford, 2015). Standard anchor test designs are ineffective in evaluating mode effects because the anchor is likely affected by mode in the same way as the total score. We found at least two studies that used background variables to evaluate test mode effects (paper versus computer) in the K–12 (Oh et al., 2015) and licensure (Puhan et al., 2007) testing contexts. In this study, we attempted

to use a similar statistical approach to adjust for group differences in ability across the TC and RP testing modes. If we could statistically remove performance differences between the two groups due to ability, then we could compare the two modes to evaluate any potential performance differences due to testing mode.

In addition to evaluating overall test performance across modes, testing programs might be interested in assessing whether the relative difficulties of test items remain constant across modes. To accomplish this, many programs use some variation of linear delta equating (Holland & Thayer, 1985; Moses, 2017). Delta equating aligns item difficulty statistics (probit-transformed proportion correct) across two groups, removing any systematic differences in item difficulty regardless of the source. Thus, if group ability and testing mode resulted in systematic differences in item statistics across the two groups, delta equating would negate all those differences regardless of the source. Even so, if only a few or some items performed differently as a function of test mode, then item-level comparisons using equated delta statistics would detect those few items displaying differential effects across TC and at-home testing.

PEG adjustment operates on differences between the two groups of examinees, irrespective of testing mode. Thus, we would expect that PEG would remove only group differences related to examinees (e.g., ability differences) while leaving differences across testing mode intact. Ideally, then, by comparing results from delta equating to PEG adjustment, we could isolate any test mode effects. In reality, the situation is more nuanced, for at least two reasons. First, the PEG approach may not remove all between-groups ability differences (Kim, 2020). Second, differential performance across TC and RP modes may result not only from mode effects but from other influences such as item exposure or test security breaches (Cohen & Wollack, 2006). Thus, unless alternative explanations can be ruled out, results of such comparisons can be ambiguous. Still, the results could provide some insight into differences across test modes.

In this study, we illustrate how to use existing test data and covariates or background information to evaluate differences in performance under RP versus TC test delivery when a randomized trial is infeasible. Using the operational data from three test titles, we applied the PEG approach to adjust for group differences in ability on a common test form given under both RP and TC mode conditions. We then assessed the effect of testing mode at both the item level and the total score level. At the item level, we compared the delta linear equating method (Holland & Thayer, 1985; Moses, 2017) to PEG adjustment to assess each item's performance across the TC and RP testing modes. At the test score level, we used the PEG approach to produce a score conversion (called RP-PEG conversion) by applying the direct equipercentile method to link the weighted (via MDIA) RP group scores to the TC group scores. Then we compared the resulting RP conversion to the TC conversion (i.e., the criterion) to determine how the RP conversion would change the RP examinees' scores and their pass/fail designations. Lastly, we proposed some requirements associated with PEG as a methodology for assessing any mode effects of RP testing.

Method

Data

We used the operational data collected in September 2020 from three licensure tests. During that period, both TC and RP examinees received the same form for each test. Tests 1 and 2 included 56 multiple-choice (MC) items. Test 3 included 40 MC items and 2 constructed response (CR) items scored by both human and machine raters using a 0 – to – 6 rating scale. The correlation between MC and CR sections was slightly higher in the RP group ($r = .39$) than in the TC group ($r = .35$). Even so, there was no significant interaction between item type and test mode ($p = .773$). Table 1 shows the descriptive statistics associated with TC and RP groups for each test. The RP group outperformed the TC group in Tests 1 and 3. There was almost no difference between the two groups on Test 2.

Procedure

Item-Level Analysis

For each test, we applied the following steps:

1. Obtain the observed delta statistics (Holland & Thayer, 1985; see also Appendix A) associated with all MC items for each of the TC and RP groups. We refer to the observed deltas associated with the TC group using the label "TC,"[3] and "RP-Observed" indicates the observed deltas associated with the RP group.

2. Obtain equated delta item statistics for the RP group (labeled "RP-Equated") via linear delta equating to TC item statistics; the equating removed overall group differences in the means and standard deviations of delta statistics (see Moses, 2017 for details; see also Appendix A).
3. Apply the weights derived from MDIA (see Haberman, 2014, for details; see also Appendix B) to RP examinees.[4] Obtain weighted delta for the RP group (labeled "RP-PEG").[5]
4. Compare delta statistics using difference (or standardized difference) plots:
    - TC versus RP-Observed (no adjustment for ability differences between the two mode groups)
    - TC versus RP-PEG (adjustment for ability differences via MDIA)
    - TC versus RP-Equated (adjustment for all systematic differences between the two groups, including any ability differences or differences related to test mode)
    - RP-Equated versus RP-PEG (may indicate differences due to mode effects).

Total Score-Level Analysis

Suppose the weights derived from the MDIA process properly adjust for group ability differences and no mode effects exist. In that case, the RP conversion via PEG linking should be nearly the same as the original TC conversion (i.e., the criterion) because both TC and RP groups took the same form of the test. Given this strong assumption, we linked the scores on the test form across TC and RP groups using direct equipercentile linking, after weighting the RP group examinees using the weights from MDIA (the same weights used for the item-level analyses). In the following analyses, we treated the TC group as the reference group and the RP group as the new group. We also treated the TC conversion as the criterion for comparisons with the resulting RP conversion. If the resulting RP conversion (RP-PEG) were close to the TC conversion, we might assume that test mode effects associated with RP testing would be negligible and that there would be no concern with using the same raw-to-scaled score conversion for both modalities.

For each test in Table 1, we applied the following steps:

1. Link the scores on the RP and TC forms using the direct equipercentile method after MDIA adjustment (RP-PEG).
2. Obtain a raw-to-scaled score conversion for the RP-PEG method.
3. Compute the unrounded scaled score difference at each raw score value between the RP-PEG conversion (from Step 2) and the TC conversion, and quantify the overall difference using the root mean squared difference (RMSD; see Equation 1).
4. Apply the conversion in Step 2 to each examinee in the RP group.
5. Compute the rounded scale score differences between the RP-PEG conversion and the TC conversion for each RP examinee (rounded RP-PEG minus rounded TC).
6. Determine each RP examinee's pass/fail designation based upon the rounded scaled scores derived in Step 4.

Table 1 Descriptive Statistics of Raw Multiple-Choice/Constructed Response/Raw Weighted Composite Scores for Test Center and Remote Proctoring Groups

Test	Group	*N*	*M*	*SD*	Percentile (5th – 95th)	Percentile (1st – 99th)
Test 1 – MC only	TC	734	31.98	8.82	17 – 46	12 – 49
	RP	1, 383	32.80	8.59	18 – 45	13 – 48
Test 2 – MC only	TC	862	34.15	9.52	17 – 48	10 – 49
	RP	1, 640	34.21	9.21	17 – 48	12 – 50
Test 3 – MC	TC	1, 003	20.38	5.40	12 – 30	9 – 33
	RP	1, 896	20.83	5.47	12 – 30	9 – 33
Test 3 – CR	TC	1, 003	13.50	2.85	8 – 17	4 – 19
	RP	1, 896	13.89	2.83	10 – 18	6 – 20
Test 3 – RWC	TC	1, 003	40.75	7.98	28 – 54	21 – 58
	RP	1, 896	41.79	8.12	29 – 55	22 – 60

Note. All three tests include six pretest items. Therefore, the maximum multiple-choice scores for Tests 1, 2, and 3 were 50, 50, and 34, respectively. For Test 3, the maximum raw weighted composite (RWC) score was 70 because of the weight of 1.5 on the CR portion (RWC = MC + 1.5 × CR). MC = multiple choice; CR = constructed response; RWC = raw weighted composite.

We computed RMSD using the formula

$$\text{RMSD} = \sqrt{\sum_{i=0}^{\max \text{RS}} f_i \left[\text{RP}_i\left(x_i\right) - \text{TC}_i\left(x_i\right)\right]^2}, \tag{1}$$

where i represents a raw score point, max RS indicates the maximum possible raw score on each test, $\text{RP}_i\left(x_i\right)$ is the unrounded scaled score for the RP-PEG conversion at raw score x, $\text{TC}_i\left(x_i\right)$ is the unrounded scaled score for the TC conversion (i.e., the criterion) at raw score x, and f_i is the relative proportion of examinees in the RP group at each raw score point.

We further estimated bias and standard errors of equating (SEE) using a resampling technique. We obtained 500 bootstrap samples (i.e., 500 replications) using the SAS PROC SURVEYSELECT procedure. In each replication, we randomly drew examinees *with replacement* from each of the TC and RP groups to match the sample sizes of the original TC and RP groups. For each replication, we obtained a new set of weights via MDIA and repeated the RP-PEG linking using the newly weighted RP sample.

Equating bias was defined as the mean difference between the RP conversion and the TC (criterion) conversion over 500 replications. The standard deviation of these differences at each score point over 500 replications was used as a measure of the conditional standard error of equating (CSEE) or error due to sampling variability. The sum of the squared bias and squared CSEE was considered an indication of total equating error variance at each score point, and the square root of this value defined the conditional root mean squared error (RMSE) index. The following equations represent bias, equating error (CSEE), and RMSE measures conditioned on each raw score point (x_i):

$$\text{Bias}_i = \bar{d}_i = \frac{\sum_{j=1}^{J}\left[\text{RP}_j\left(x_i\right) - \text{TC}\left(x_i\right)\right]}{J}, \tag{2}$$

$$\text{CSEE}_i = sd_i = \sqrt{\text{Var}_i\left[\text{RP}_j\left(x_i\right) - \text{TC}\left(x_i\right)\right]} = \sqrt{\text{Var}_i\left[RP_j\left(x_i\right)\right]}, \tag{3}$$

$$\text{RMSE}_i = \sqrt{\bar{d}_i^2 + sd_i^2}, \tag{4}$$

where j is a replication, J is the total number of replications (500), Var_i is the variance of the differences between the RP and TC scaled scores across 500 samples at each raw score point, and $\text{RP}_j(x)$ denotes the scaled score equivalent calculated from a PEG linking method in the sample j.

As overall summary measures, we computed the weighted average root mean squared bias, $\sqrt{\sum_i f_i \text{Bias}_i^2}$, the weighted average standard error of equating, $\sqrt{\sum_i f_i \text{SEE}_i^2}$, and the weighted average RMSE, $\sqrt{\sum_i f_i \text{RMSE}_i^2}$, across the RP group score distribution, where f_i is the relative proportion at each score point in the raw score distribution of the RP group.

Results

Item-Level Analysis

The TC group has a single set of item delta statistics, which is the observed TC deltas. In contrast, the RP group has three sets of deltas: observed (RP-Observed), linearly equated (RP-Equated), and weighted via MDIA (RP-PEG). Table 2 presents the summary statistics associated with those four sets of deltas. Owing to linear equating, the means and SDs of the RP-Equated deltas were identical to those of deltas in the TC group (TC). For Tests 1 and 3, RP-Observed showed lower delta means than did TC, indicating the TC group's inferior performance. Suppose PEG perfectly adjusted for group differences in ability. Then any remaining differences in delta values between TC (or RP-Equated) and RP-PEG would indicate differences across mode, including mode effects and any differential security incidents across mode (e.g., item exposure, cheating). Assuming successful adjustment for most or all ability differences, the magnitude of mode differences seems generally small (favoring the RP group) for Tests 1 and 3 on average. For Test 2, no differences emerged among the four sets of deltas, implying no mode effects and no group ability differences between the TC and RP groups.

Table 2 Descriptive Statistics Associated With Four Sets of Multiple-Choice Item Deltas in Each Test

Test	*N*	Delta	*M*	*SD*	Min	Max
1	56	TC	11.55	1.98	6.34	15.61
		RP-Observed	11.32	2.10	5.65	15.38
		RP-PEG	11.44	2.09	5.72	15.67
		RP-Equated	11.55	1.98	6.19	15.39
2	56	TC	11.00	1.79	7.56	15.59
		RP-Observed	10.98	1.81	7.64	15.74
		RP-PEG	10.98	1.80	7.67	15.71
		RP-Equated	11.00	1.79	7.68	15.71
3	40	TC	11.74	1.98	6.38	15.24
		RP-Observed	11.58	2.06	5.75	14.84
		RP-PEG	11.65	2.05	5.83	14.95
		RP-Equated	11.74	1.98	6.15	14.87

Note. N = number of multiple-choice items (operational plus pretest items); TC = test center; RP = remote proctoring; PEG = pseudo-equivalent groups.

Figure 1 compares item deltas across the TC and RP groups. The three columns in Figure 1 show results for Tests 1, 2, and 3, respectively. In each column, the three plots present the difference between the observed delta set in the TC group (TC) and each of the three delta sets in the RP group: RP-Observed (top), RP-PEG (middle), and RP-Equated (bottom). For Test 2, differences between TC and RP-Observed deltas were centered around zero, indicating no general tendency for the items to be easier or harder for the RP group than for the TC group. For Tests 1 and 3, however, most differences for RP-Observed deltas were located below the zero line, showing that the RP group found the items easier. Even so, most delta differences were smaller than 1.0 in absolute value. For all tests, PEG (middle row) and delta equating (bottom row) resulted in reasonably comparable delta values in terms of patterns and magnitudes. Test 1 involved the largest adjustment via those two methods, but Test 2 item statistics showed almost no adjustment.

The three plots in Figure 2 graphically display the delta differences between RP-PEG and RP-Equated in the RP group. More specifically, those plots indicate whether the two methods (PEG versus delta equating) yielded inconsistent results (e.g., an item looked like an outlier by one method but not by the other), which would indicate the possible presence of test mode effects. For Tests 1 and 3, the PEG deltas were lower than equated deltas, indicating some systematic differences. Even so, the differences in delta values between delta equating and PEG in the RP group were minimal and smaller than 0.5 in absolute value.

Table 3 summarizes how many items appeared to be functioning differently across the TC and RP groups when either delta equating or the PEG approach was applied to adjust for group differences. An item was flagged for differential functioning if its standardized difference exceeded 2.5 in absolute value. One can confirm this information against the standardized difference plots presented in Figures C1 – C3 in Appendix C. In the standardized difference plots, the dotted lines at ±2.5 show the cutoff criterion for specifying differentially functioning items across the TC and RP groups.[6] The PEG result agreed closely enough with delta equating in flagging problematic outliers. Overall, the differential item-level effects related to mode of testing seem to be only minor for all three tests.

Total Score-Level Analysis

Table 4 presents the descriptive statistics associated with the TC and RP groups on each test. We computed the RP group's means and SDs based on the two sets of rounded scaled scores derived from the TC and RP-PEG conversions. The RP-PEG score is the scaled score the RP examinee would have received if the RP-PEG conversion had been used to score the RP examinee. As with the item-level analyses, Test 2 exhibited no differential effects related to mode. For Tests 1 and 3, the mean scaled score from the RP-PEG conversion was lower than that derived from the TC conversion, indicating some score-level differences across mode even after PEG adjustment. Even so, the magnitude of the differences was minimal.

The three plots in Figure 3 present the conditional unrounded scaled score differences across the raw score range where most RP examinees were located (from the 1st to 99th percentiles; see Table 1) for each of the three tests. In each plot, the solid horizontal line at zero represents identical values across the RP-PEG conversion and the TC conversion. The two dotted horizontal lines at ±2.0 indicate the interval between two adjacent possible scaled scores (for all three tests, only

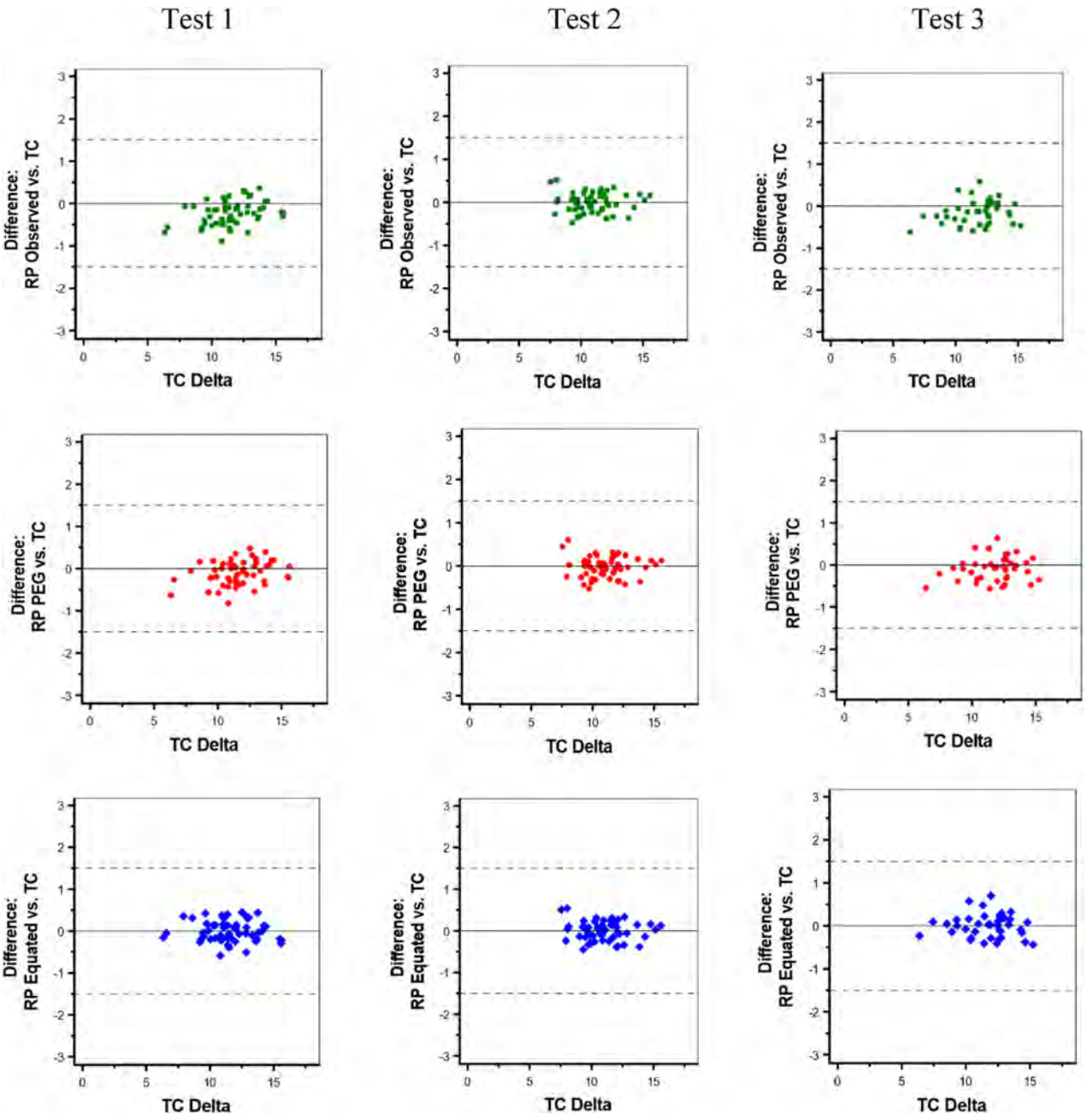

Figure 1 Difference plots between TC delta and each of three RP delta sets in Tests 1, 2, and 3. *Note*. Difference = RP minus TC; TC = test center; RP = remote proctoring.

even scaled scores are awarded). The solid red curve depicts the conditional scaled score differences between the RP-PEG conversion and the TC conversion at each raw score. The solid vertical line indicates the cut score on the test.[7] For all three tests, the unrounded RP-PEG conversion was lower than the TC conversion across most, if not all, of the raw score range, but most importantly at the cut score. The unrounded difference at the cut score (using the TC conversion) was most pronounced for Test 1.

For all three tests, examinees receive rounded scaled scores, ranging from 100 to 200 with an interval of 2. We calculated the number of RP examinees whose scaled scores differed when using different conversions (TC or RP-PEG conversion) and by how much. Table 5 presents distributions of the scaled score differences between the RP-PEG conversion and the TC conversion computed from the RP group examinees in each test. For Test 2, the rounded scaled scores remained

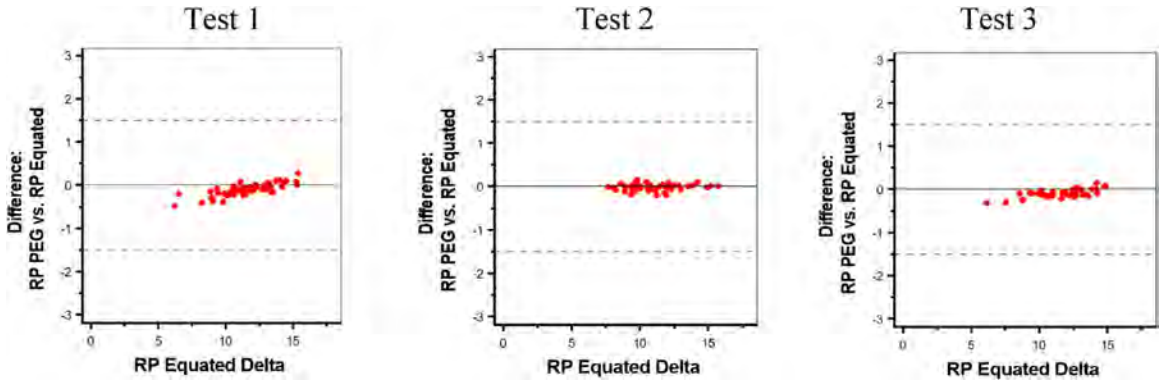

Figure 2 Delta difference plots between RP-PEG and RP-Equated in each test. *Note*. Difference = RP-PEG minus RP-Equated; RP = remote proctoring; PEG = pseudo-equivalent groups.

Table 3 The Number of Differentially Functioning Items Across Test Center and Remote Proctoring Conditions Detected by Delta Equating and Pseudo-Equivalent Groups in Each Test

Test	N of MC items	Delta equating	PEG	Note for flagged items
1	56	0	1	PEG flagged an item; harder in TC (TC P+ = 0.71; RP-PEG P+ = 0.77)
2	56	0	1	PEG flagged an item; harder in RP (TC P+ = 0.89; RP-PEG P+ = 0.86)
3	40	1	0	Delta flagged an item; harder in RP (TC P+ = 0.6; RP P+ = 0.53)

Note. The item flagging decision was made upon the cutoff criterion of ±2.5 on the standardized difference scale, which is the current practice of at least one large-scale testing program. MC = multiple choice; TC = test center; RP = remote proctoring; PEG = pseudo-equivalent groups.

Table 4 Descriptive Statistics of Rounded Scaled Scores Derived From Two Different Conversions for the Test Center and Remote Proctoring Groups in Each Test

Test	Group	Conversion	*N*	*M*	*SD*	Percentile (1st–99th)
1	TC	TC	734	169.8	22.2	118–200
	RP	TC	1,383	171.9	21.7	120–200
	RP	RP-PEG	1,383	170.4	22.0	116–200
2	TC	TC	862	163.7	24.3	100–200
	RP	TC	1,640	163.8	23.6	106–200
	RP	RP-PEG	1,640	163.5	24.4	100–200
3	TC	TC	1,003	158.2	13.4	130–188
	RP	TC	1,896	160.0	13.7	132–192
	RP	RP-PEG	1,896	158.7	13.6	130–190

Note. TC = test center; RP = remote proctoring; PEG = pseudo-equivalent groups.

unchanged for 73% of the RP examinees. The proportion of RP examinees whose scores remained the same was much smaller in Tests 1 and 3. For those tests, approximately two thirds of RP examinees' scores decreased by 2 points when the RP-PEG conversion was applied. Even so, the differences were primarily within the range of −2 to +2 after rounding. Note that all three tests use a score interval of 2.

We also examined the RP examinees' pass/fail designations based upon their two sets of scaled scores derived from the TC and RP-PEG conversions. As presented in Table 6, the pass/fail status of most RP examinees remained the same regardless of which conversion was used to report their scores. After rounding, the RP-PEG cut scores were the same as the TC cut scores for Tests 1 and 2. This means that an examinee scoring at the indicated cut score would pass the test whether the RP-PEG conversion or the TC conversion was used to score the test. However, this was not the case for Test 3. The rounded RP-PEG conversion was lower than the rounded TC conversion at the cut score, leading to some inconsistency in pass/fail status. The status of about 6% of the RP examinees would have changed from pass to fail if the RP-PEG conversion had been used.[8] No one's status would change from fail to pass.

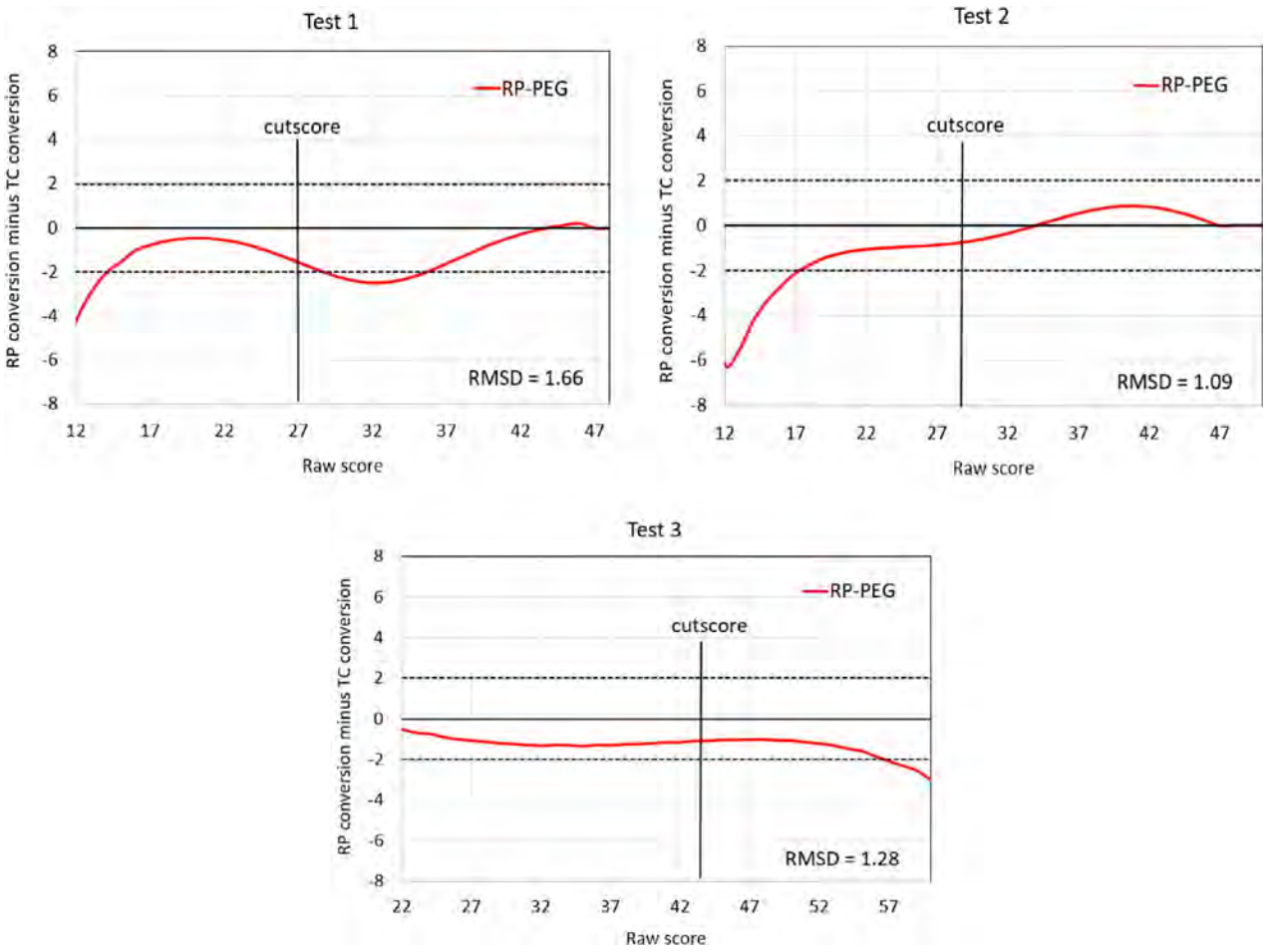

Figure 3 Unrounded scaled score differences between the remote proctoring – pseudo-equivalent groups conversion and the test center conversion in Tests 1, 2, and 3. *Note*. For each test, all stakeholders use the same cut score. TC = test center; RP = remote proctoring; PEG = pseudo-equivalent groups.

Table 5 Distribution of Differences Between the Scaled Scores Derived From the Test Center Conversion and Those Derived From the Remote Proctoring – Pseudo-Equivalent Groups Conversion in the Remote Proctoring Group

	Differences from TC conversion				
Test	Below −2	−2	0	+2	Above +2
1 (n = 1,383)	132 (9.5%)	800 (57.9%)	451 (32.6%)	—	—
2 (n = 1,640)	46 (2.8%)	273 (16.7%)	1,194 (72.8%)	127 (7.7%)	—
3 (n = 1,896)	10 (0.5%)	1,194 (63.0%)	687 (36.2%)	2 (0.1%)	3 (0.2%)

Note. All tests use a score interval of 2. TC = test center.

Table 7 summarizes the weighted average root mean squared bias, equating error, and RMSE derived from a bootstrap resampling technique for the RP-PEG conversion. Test 3, having larger samples in size, showed smaller SEE than did Tests 1 and 2. Figure 4 displays plots of the RP-PEG conversion's conditional bias along with an error band representing plus or minus two empirical CSEEs (95% error band). Although the conditional bias was nontrivial across the raw score points, the 95% error bands include the zero line (no difference from the TC conversion) mainly because of wide error bands. This was particularly true for Tests 1 and 2. For Test 3, the error band was much tighter than the one for Tests 1 and 2. Therefore, any systematic differences between the RP conversion and the TC conversion could be considered as the result of sampling fluctuation. For Test 3, the 6% pass/fail inconsistency for the RP examinees could be regarded as sampling variability.

Table 6 RP Examinees' Pass/Fail Designations Associated With Test Center and Remote-Proctoring Pseudo-Equilvalent Groups Conversions in Each Test

Test	TC conversion	RP-PEG conversion	
		Fail	Pass
1	Fail	344 (24.9%)	—
	Pass	—	1,039 (75.1%)
2	Fail	438 (26.7%)	—
	Pass	—	1,202 (73.3%)
3	Fail	1,016 (53.6%)	—
	Pass	110 (5.8%)	770 (40.6%)

Note. TC = test center; RP = remote proctoring; PEG = pseudo-equivalent groups.

Table 7 Summary of Bootstrapped Weighted Average Root Mean Squared Bias, Equating Error, and Root Mean Squared Error Associated With the Remote Proctoring – Pseudo-Equivalent Groups Conversion in Each Test

Test	RP-PEG conversion		
	Bias	SEE	RMSE
1	1.19	1.36	1.80
2	1.15	1.34	1.76
3	1.27	0.77	1.49

Note. SEE = standard error of equating; RP = remote proctoring; PEG = pseudo-equivalent groups; RMSE = root mean squared error.

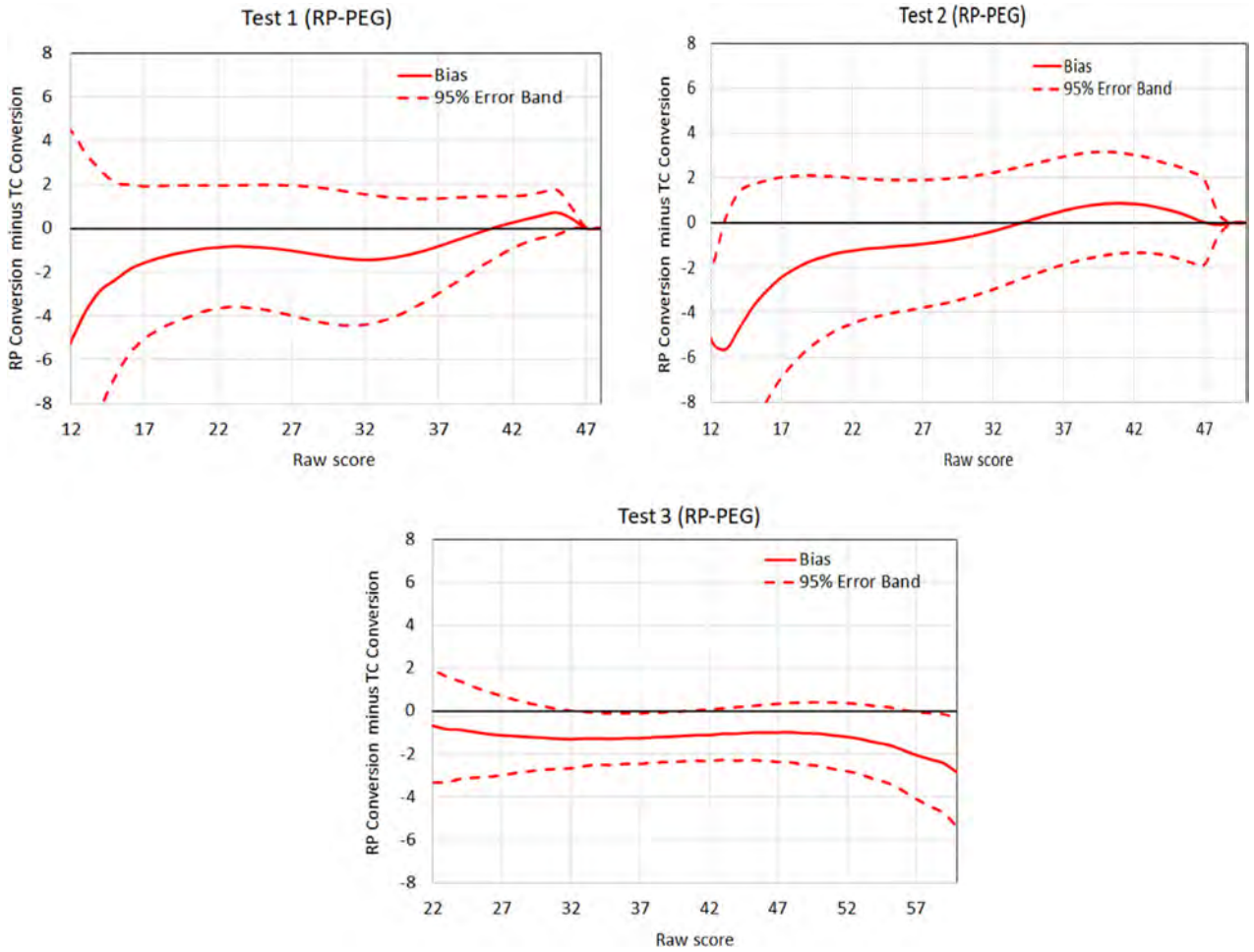

Figure 4 Differences of the remote proctoring – pseudo-equivalent groups conversion from the test center conversion and its 95% error band in Tests 1, 2, and 3. *Note*. TC = test center; RP = remote proctoring; PEG = pseudo-equivalent groups.

Discussion

As a response to the global pandemic, many testing programs have added RP at-home testing as an available testing mode. The score comparability between examinees testing remotely and those testing in a TC is critical, because test scores have significant consequences for examinees and other score users. Whenever testing programs offer multiple test modes (e.g., paper, computer, onsite testing, remote testing) to expand their service, they need to be mindful of the possible interaction of those modes with the content or constructs measured by a test, which could impede fair assessment of examinee ability. Any expansion of modes of testing must maintain fairness to all examinees. Therefore, testing programs need to gather substantial amounts of psychometric evidence to support mode comparability.

Psychometricians on testing programs routinely examine item statistics to determine if the relative difficulty of certain items has shifted from one administration to the next. A commonly used method (delta equating) involves equating the item statistics from a new sample to match as closely as possible the statistics in a previous sample. Then, psychometricians examine item difficulty differences across samples for large deviations. These deviations signal possible problem items. Delta equating directly adjusts for all systematic differences in item performance across two groups (traditional delta equating is nonspecific), including both group ability differences and mode effects. Therefore, if the difference in the mode of testing affects all (or most) of the items, traditional delta equating will not reveal that effect. This is a clear limitation of traditional delta equating as a method to detect mode effects of testing. In this study, we used the PEG approach as an alternative to resolve this limitation.

The PEG methodology uses background information concerning examinees to construct sample weights via MDIA. The sample weights transform nonequivalent groups of examinees into groups that resemble equivalent groups on selected background variables, potentially removing differences in ability as well. Therefore, delta plots based on PEG weighting would adjust for performance differences based on a specific source, namely group differences. Thus, if the PEG assumption holds (i.e., the weights properly adjust for group ability differences), mode effects would be a plausible explanation for any differences between PEG and delta equating of RP to TC test items. By comparing PEG delta values to equated delta values, we may be able to isolate systematic performance differences due to mode effects or test exposure.

This study illustrated a potential means for evaluating mode effects of RP testing without a randomized trial. For this purpose, we used actual operational data from three licensure tests administered in September 2020. Using the PEG methodology, we were able to separate ability differences between groups testing under TC and RP conditions from other differential effects related to mode. We assumed that PEG adjusts for group ability differences between TC and RP reasonably well, and we assumed the absence of any test security issues, so that we could attribute any systematic differences between TC and RP at the item level and the total score level to test mode effects. For Test 2, the differences in item statistics between PEG and delta equating were not only trivial but also nonsystematic. The differences between the RP group conversion and the TC group conversion were also minor at the total score level, leading to consistent pass/fail designations for all RP examinees. For Tests 1 and 3, the differences between RP and TC groups were small at both the item and total score levels, but the differences led to different scaled scores or inconsistent pass/fail decisions (this was the case for Test 3) for approximately two thirds of the RP examinees. Thus, although small, performance differences across test mode were consequential.

In general, the PEG methodology requires two conditions: one is a sufficiently large sample for each test mode group, and another is the existence of adequate demographic and background variables. PEG is essentially a poststratification procedure that uses demographic variables to create statistically matched groups of examinees. The success of this approach depends on the relevance of the available demographic information. If the demographics are closely related to the test score, the PEG adjustment will remove most of the ability difference, leading to reasonably equivalent adjusted groups. In that case, any remaining differential performance between groups could be attributed to differences across test modes.

Note that in practice, the performance of PEG is rarely testable. We can reasonably assume that the stronger the relationship between the covariates used in the MDIA procedure and examinees' performance on the test, the better the PEG approach will be at adjusting for group differences in ability. As shown in Table C1, the R^2 values for the regression of test scores on covariates ranged from .24 to .39 (adjusted $R^2 = .21 – .33$) over the six regression models constructed as a function of two test mode groups and three test titles. Given the moderate size of these correlations, we might expect that the PEG approach only partially adjusted for group ability differences in this study.

We could produce a more direct test of the effectiveness of PEG at adjusting for group ability differences by examining different groups taking the same test under the same testing mode. In this situation, we could reasonably attribute any

performance differences across groups to ability differences. Previous research (Kim, 2020) with tests and covariates like the ones used here found that PEG generally eliminated most, but not all, group differences. Thus, it would be more legitimate to regard the differences between TC and PEG-adjusted RP results as a mix of unadjusted ability differences and some other systematic factors (e.g., test mode effects or a security breach).

The degree of departure of the adjusted RP from the TC conversion, taken together with PEG's limitation for adjusting these groups in ability, may indicate minor test mode effects at the total score level. Overall, the present findings offered some empirical evidence for small mode effects associated with the RP testing for the tests examined in this study. As mentioned before, alternate scenarios could explain our results. For example, the greater overall performance of the RP group could be a combination of higher ability (partially corrected by the PEG approach), some security issue boosting RP scores, and mode effects that suppress RP scores, resulting in a small net RP advantage after PEG adjustment. In the absence of more definitive evidence, we will cautiously adopt the more parsimonious view of small mode effects.

For tests used in this study, more examinees preferred the RP mode over the TC mode. This trend is common for many tests being offered via the RP mode of testing. It would be interesting to examine whether there exist systematic differences or salient patterns in terms of mode group memberships (e.g., rural examinees prefer RP over TC; repeaters prefer RP over TC, whereas first timers prefer TC over RP). If so, those differences should be utilized in the process of minimizing discriminant information between the two mode groups.

In general, examinees' previous performance levels can be considered a strong indicator for predicting their present performance levels. Thus, it may be worthwhile to apply this methodology to repeat examinees to assess if there exist systematic patterns or interactions between repeaters' performance change and their choice on testing mode.[9] Even so, we should not rely solely on repeaters to make judgments about adjustments for the total population without demonstrated proof of the invariance of linking functions across the populations of repeaters and first-time examinees. In any event, as mentioned previously, PEG is not a perfect solution for assessing test mode effects. However, it does show promise as a convenient and cost-effective diagnostic tool.

The introduction of at-home testing raises concerns such as test mode effects and test security. Some conventional individual cheating behaviors, such as copying answers from other examinees, are virtually impossible when the test taker is testing alone and being monitored and videotaped as in the RP administration. Even so, other types of cheating behaviors, such as keeping notes or unauthorized materials in a room, having someone else take the test, or communication with outsiders, could possibly occur. In this study, however, we did not consider potential impact of security breaches in the process of adjusting for group differences between TC and RP testing groups in ability. Further investigation will be needed to examine the impact of security breaches such as cheating and item exposure on the performance of RP examinees.

Notes

1 Pommerich (2016) pointed out that comparability is not the same as equivalence, which we might expect if two sets of scores are equated across two forms of the same test meeting certain requirements. Comparability is a weaker form of comparison in which we can identify concordant score points across the two forms, but not strict equivalence in terms of measured constructs and predictive validity.

2 In the context of equating, we would generally expect the relationship of test scores to background information (e.g., demographic information or available measures of constructs similar to that measured by the test) to be much weaker than the relationship of test scores to anchor scores (usually shorter measures of the same construct measured by the test).

3 Because only observed delta statistics were computed from the TC group, we refer to TC group's observed delta statistics simply as "TC" in this study.

4 A total of 68 dummy variables constructed from 18 background questions (e.g., gender, age, education level, undergraduate major, GPA, region, repeater status) were used in the process of MDIA. In Appendix C, Table C1 presents the *R*-squared (together with adjusted *R*-squared) values derived from regression analysis in which dummy-coded demographic variables were used to predict the examinees' scaled scores in each of the TC and RP groups.

5 We can offer the SAS code upon request.

6 In the standardized difference plot, the dotted lines at ±2.5 indicate a cutoff criterion for detecting outliers in a distribution—observations (i.e., standardized differences) that are so different from the others that we do not believe they belong in the distribution. The choice of the value 2.5 is arbitrary. If the distribution is approximately normal (0.1), a cutoff value of 2.5 means that we will not label an observation as an outlier unless it has a very small probability (about .01) of occurring, if it really

belongs in the distribution. If there are no real differences, that is, if the items all really belong in the distribution, we will identify very few of items as outliers (only about 1% of them).

7 The vertical cut score line in the figure is drawn at the raw score corresponding to the scaled cut score using the TC conversion. For each test, all stakeholders use the same cut score.

8 As presented in Table 1, Test 3 is a mix of MC and CR items. The RP group's overall performance on MC and CR sections remained the same regardless of whether the RP examinees were weighted via MDIA or not weighted. The MC and CR means of the weighted RP group were 20.62 and 18.83, respectively. Those values were comparable to the MC and CR means of the unweighted RP group (see Table 1). There was no interaction between item type (MC/CR) and test mode (TC/RP) before and after weighting.

9 Because the data sets we have include only examinees' repeater status but not their previous scores, we were not able to assess this matter in this study.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Cohen, A. S., & Wollack, J. A. (2006). Test administration, security, scoring, and reporting. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 355–386). Praeger.

Eignor, D. R. (2007). Linking scores derived under different modes of test administration. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 135–159). Springer. https://doi.org/10.1007/978-0-387-49771-6_8

Haberman, S. J. (2014). *A program for adjustment by minimum discriminant information* (Research Memorandum No. RM-14-01). ETS.

Haberman, S. J. (2015). Pseudo-equivalent groups and linking. *Journal of Educational and Behavioral Statistics*, *40*(3), 254–273. https://doi.org/10.3102/1076998615574772

Holland, P. W., & Thayer, D. T. (1985). *An alternate definition of the ETS delta scale of item difficulty* (Research Report No. RR-85-43). ETS. https://doi.org/10.1002/j.2330-8516.1985.tb00128.x

Kim, S. (2020, August 5). *The effectiveness of the pseudo-equivalent groups (PEG) procedure in statistically adjusting for group differences in certification tests* [Unpublished manuscript]. ETS.

Kim, S., & Lu, R. (2018). *The pseudo-equivalent groups approach as an alternative to common-item equating* (Research Report No. RR-18-02). ETS. https://doi.org/10.1002/ets2.12195

Livingston, S. A. (2014). *Demographically adjusted groups for equating test scores* (Research Report No. RR-14-30). ETS. https://doi.org/10.1002/ets2.12030

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of equating and sampling works best? *Applied Measurement in Education*, *3*(1), 73–95. https://doi.org/10.1207/s15324818ame0301_6

Longford, N. T. (2015). Equating without an anchor for nonequivalent groups of examinees. *Journal of Educational and Behavioral Statistics*, *40*(3), 227–253. https://doi.org/10.3102/1076998615574773

Lu, R., & Guo, H. (2018). *A simulation study to compare nonequivalent groups with anchor test equating and pseudo-equivalent group linking* (Research Report No. RR-18-08). ETS. https://doi.org/10.1002/ets2.12196

Mislevy, R. J, Sheehan, K. M., & Wingersky, M. (1993). How to equate tests with little or no data. *Journal of Educational Measurement*, *30*(1), 55–78. https://doi.org/10.1111/j.1745-3984.1993.tb00422.x

Moses, T. (2017). A review of developments and applications in item analysis. In R.E. Bennett & M. von Davier (Eds.), *Advancing human assessment, methodology of educational measurement and assessment* (pp. 19–46). Springer. https://doi.org/10.1007/978-3-319-58689-2_2

Oh, H., Liu, J., & Gaj, S. (2015, April 15–19). *Application of PEG linking for testing mode adjustment in K–12 assessment* [Paper presentation]. The annual meeting of the National Council on Measurement in Education, Chicago, IL, United States.

Pommerich, M. (2016). The fairness of comparing test scores across different tests or modes of administration. In N. J. Dorans & L. L. Cook (Eds.), *Fairness in educational assessment and measurement* (pp. 111–134). Routledge.

Puhan, P., Boughton, K., & Kim, S. (2007). Examining differences in examinee performance in paper and pencil and computerized testing. *Journal of Technology, Learning, and Assessment*, *6*(3). https://ejournals.bc.edu/index.php/jtla/article/view/1633

Wallin, G., & Wiberg, M. (2019). Kernel equating using propensity scores for nonequivalent groups. *Journal of Educational and Behavioral Statistics*, *44*(4), 390–414. https://doi.org/10.3102/1076998619838226

Wiberg, M., & Bränberg, K. (2015). Kernel equating under the non-equivalent groups with covariates design. *Applied Psychological Measurement*, *39*(5), 349–361. https://doi.org/10.1177/0146621614567939

Wright, N. K., & Dorans, N. J. (1993). *Using the selection variable for matching or equating* (Research Report No. RR-93-04). ETS. https://doi.org/10.1002/j.2333-8504.1993.tb01515.x

Xi, N., Guo, H., & Oh, H. (2015, April 15–19). *A PEG linking study of matching variables* [Paper presentation]. The annual meeting of the National Council on Measurement in Education, Chicago, IL, United States.

Appendix A

A Linear Delta Equating

The most well-known item difficulty index is the average item score or the proportion of correct responses (p or p+) for dichotomously scored items. At ETS, most testing programs use the difficulty of an item measured by a transformation of p to the "delta scale." P can be transformed into delta (Δ) using Equation A1 (see Holland & Thayer, 1985; Moses, 2017 for details).

$$\Delta\ \left(p\right) = 13 - 4\Phi^{-1}\left(p\right), \tag{A1}$$

where $\Phi^{-1}\left(p\right)$ denotes the inverse function of the normal cumulative distribution function.

Testing programs examine variations on item difficulty to address examinee group influences on item difficulty indices. They describe these variations by examining both actual item difficulty parameters and adjustments to existing item difficulty parameters (i.e., reference group item parameter estimates). A common adjustment is to transform the mean and standard deviation of a set of new delta values from one examinee group (e.g., new group; RP group) to this set's mean and standard deviation from another examinee group (e.g., reference group; TC group). The transformation of one group's item deltas to the scale of another group's deltas is often called delta equating, linear delta equating, or mean–sigma equating because it requires the means and standard deviations of the two sets of delta values. Also, it reflects a linear relationship between the deltas obtained from two examinee groups on the same set of items. Equation A2 shows how to transform RP group's item delta to the delta scale of TC group in order to get equated deltas of the RP group.

$$\widehat{e}_{\mathrm{TC}}\left(\widehat{\Delta}_{i,\mathrm{RP}}\right) = \overline{\Delta}_{\mathrm{TC}} + \frac{\widehat{\sigma}_{\mathrm{TC}}\left(\Delta\right)}{\widehat{\sigma}_{\mathrm{RP}}\left(\Delta\right)}\left(\widehat{\Delta}_{i,\mathrm{RP}} - \overline{\Delta}_{\mathrm{RP}}\right), \tag{A2}$$

where $\overline{\Delta}_{\mathrm{TC}}$ = the average of TC group's deltas, $\overline{\Delta}_{\mathrm{RP}}$ = the average of RP group's deltas, $\widehat{\sigma}_{\mathrm{TC}}\left(\Delta\right)$ = the standard deviation of TC group's deltas, $\widehat{\sigma}_{\mathrm{RP}}\left(\Delta\right)$ = the standard deviation of RP group's deltas, and $\widehat{\Delta}_{i,\mathrm{RP}}$ = a particular item delta in the RP group.

Appendix B

Weights to Construct Pseudo-Equivalent Groups

Assume that there are two groups (RP and TC in this study) who took the same questionnaire designed to collect examinees' background information. Let x_i be the score of Examinee i in the RP group, and y_j be the score of Examinee j in the TC group. Let z_{iX} and z_{jY} represent the K-dimensional Vector $\mathbf{Z}$ of background variables collected through the questionnaire for Examinee i on RP and for Examinee j on TC, respectively. The elements in the $\mathbf{Z}$ vector are multiple independent dummy variables coded from each of the background questions on the questionnaire. Under this condition, the target background mean is the averaged background $\mathbf{Z}$ vectors over the target group, which is the TC group in this study, as shown in Equation B1.

$$\overline{Z} = \frac{\left(\sum_{i=1}^{N_Y} z_{iY}\right)}{N_Y}, \tag{B1}$$

where N_Y is the number of examinees in the TC group.

$\overline{Z}$ is then used as the target background vector to construct PEGs. In this step, each examinee of the RP group gets a weight w_{iX} so that the weighted background vector of the RP group is matched to the target Vector $\overline{Z}$, as shown in Equation B2.

$$\frac{\left(\sum_{i=1}^{N_X} w_{ix} z_{ix}\right)}{N_X} = \overline{Z}, \tag{B2}$$

where N_X is the number of examinees in the RP group, $w_{iX} > 0$, and $\sum_{i=1}^{N_X} w_{ix} = 1$. Here the adjustment of the minimum discriminant information (MDIA) approach with the Newton – Raphson method is used to obtain the individual weight w_{iX} (see Haberman, 2014, 2015 for details). In this study, each examinee in the RP group received a weight derived from MDIA. Then, weighted item average scores (p+, delta) and weighted total test scores (either MC or RWC [raw weighted composite]) were used at the item-level analysis and the total score-level analysis, respectively. For PEG linking using direct equipercentile, RP-PEG used the weighted score distribution of the RP group. The frequencies on scores (i.e., grouped data) were counted based on those weighted examinees. The percentile ranks of the RP group were determined based on the grouped data.

Appendix C

Supplemental Analysis Results

Table C1 *R*-squared values from regression models predicting scaled scores from dummy-coded demographic variables

Test	# of dummy predictors	Test center	Remote proctoring
1	68	0.39 (0.33)	0.36 (0.33)
2	68	0.35 (0.30)	0.27 (0.24)
3	68	0.29 (0.24)	0.24 (0.21)

Note. The values in the parentheses indicate adjusted *R*-squared values from the regression models. Dummy predictors were constructed based on 18 background questions (e.g., gender, age, major, GPA, educational level, first language, repeater status).

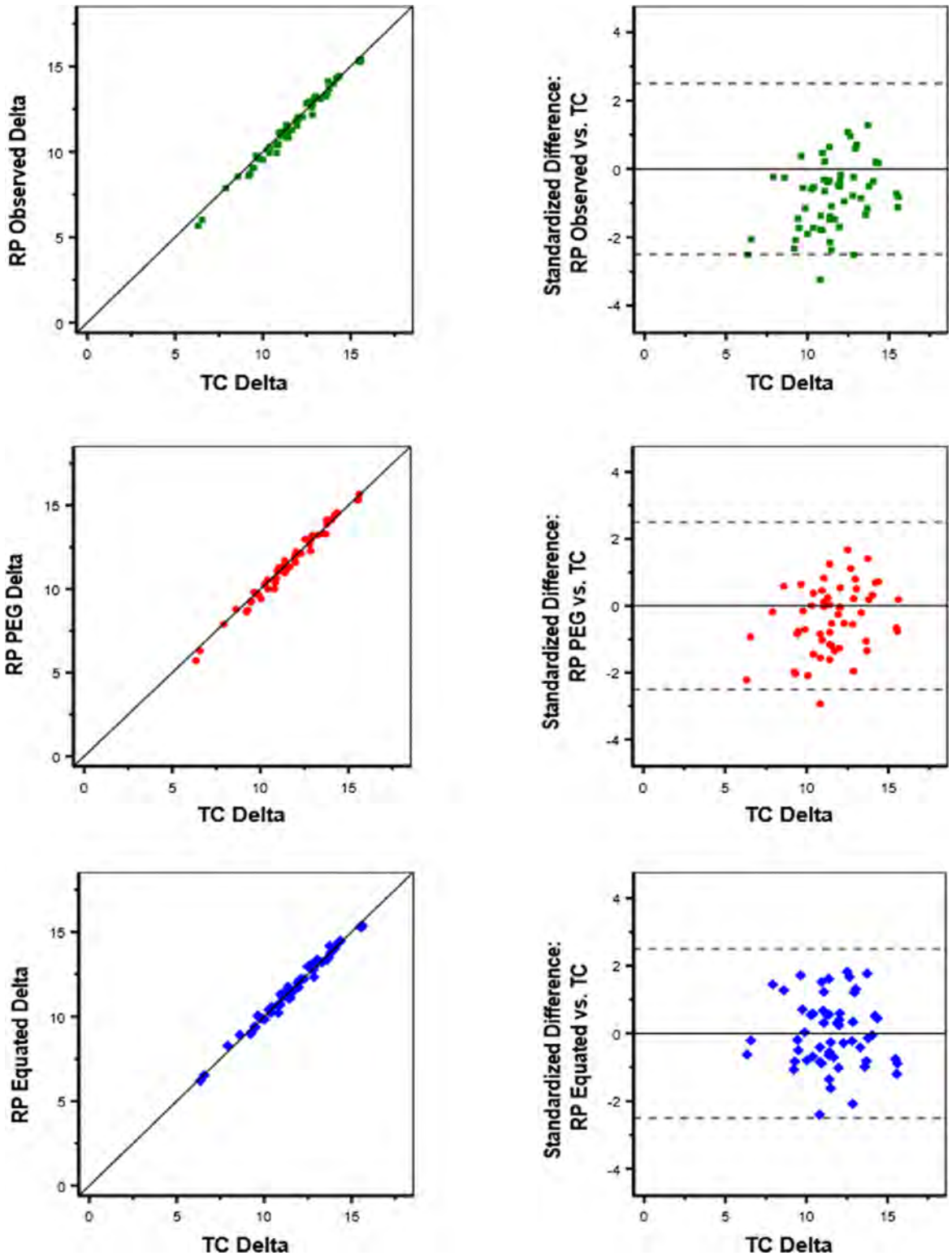

Figure C1 Deltas and standardized delta difference plots for Test 1.

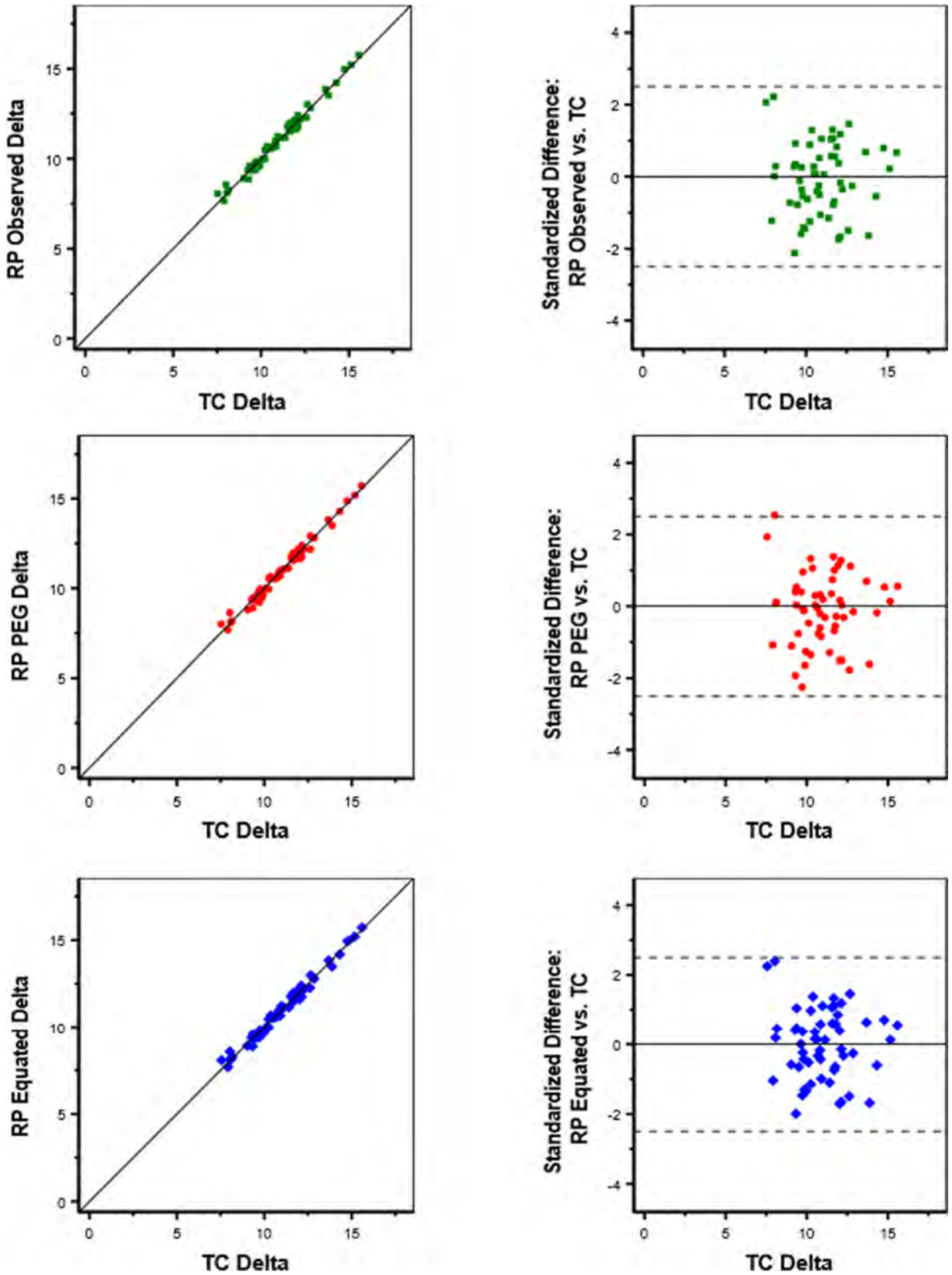

Figure C2 Deltas and standardized delta difference plots for Test 2.

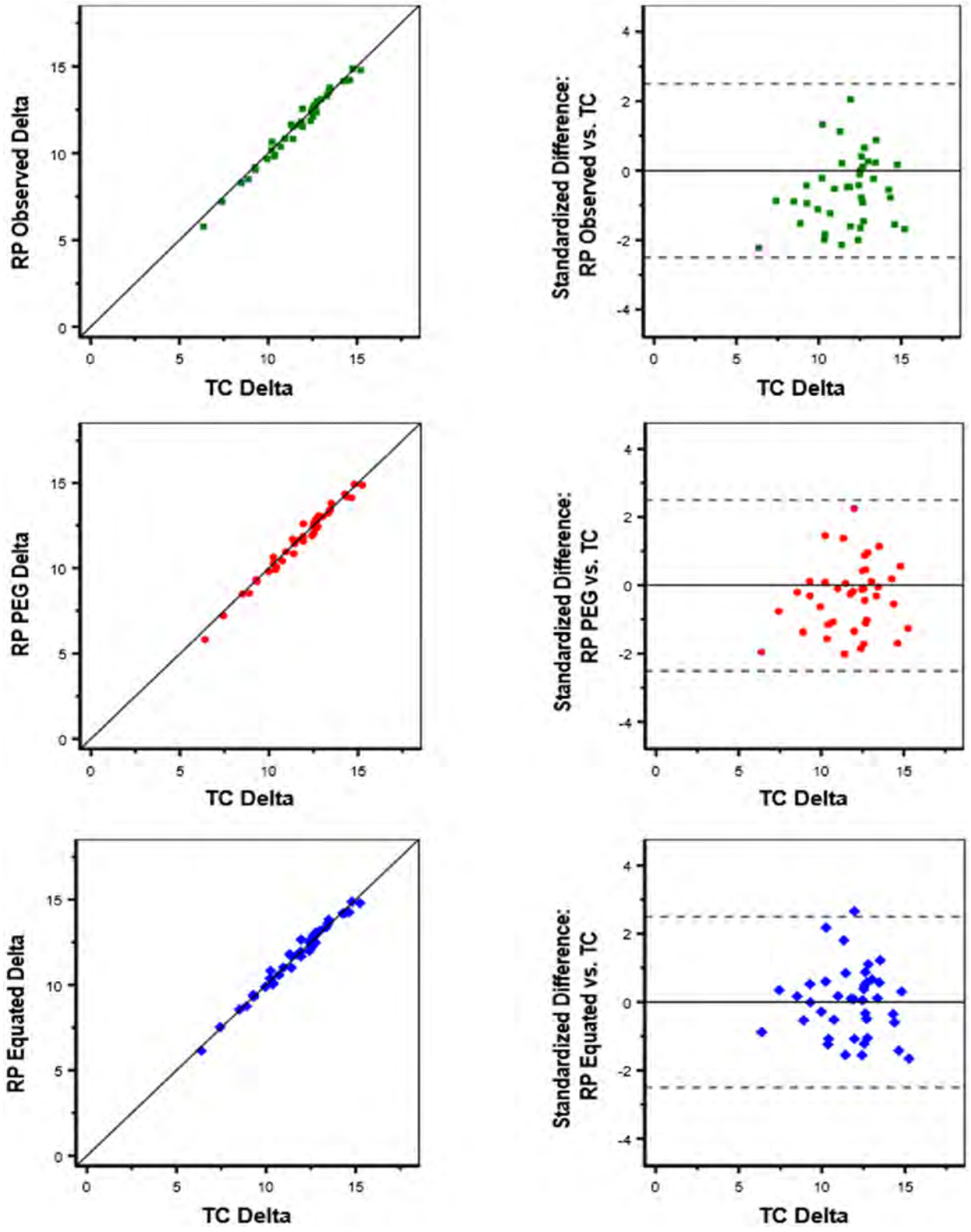

Figure C3 Deltas and standardized delta difference plots for Test 3.

Suggested citation:

Kim, S., & Walker, M. (2021). *Assessing mode effects of at-home testing without a randomized trial* (Research Report No. RR-21-10). ETS. https://doi.org/10.1002/ets2.12323

Action Editor: John Mazzeo

Reviewers: Sandip Sinharay and Ru Lu