*ets research institute

DECEMBER 2024
RR–24-11

RESEARCH REPORT

Detecting the Impact of Remote Proctored At-Home Testing Using Propensity Score Weighting

AUTHORS

Jing Miao, Yi Cao, and Michael E. Walker

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

ETS RESEARCH REPORT

Detecting the Impact of Remote Proctored At-Home Testing Using Propensity Score Weighting

Jing Miao[1], Yi Cao[1], & Michael E. Walker[2]

1 ETS Psychometric Analysis & Research, ETS, Princeton, New Jersey, United States
2 Human Resources Research Organization (HumRRO), Alexandria, Virginia, United States

Studies of test score comparability have been conducted at different stages in the history of testing to ensure that test results carry the same meaning regardless of test conditions. The expansion of at-home testing via remote proctoring sparked another round of interest. This study uses data from three licensure tests to assess potential mode effects associated with the dual option of on-site testing at test centers and at-home testing via remote proctoring. We generated propensity score weights to balance the two self-selected groups in order to detect the mode effect on the test outcomes. We also assessed the potential impact of omitted variables on the estimated mode effect. Results of the study indicate that the demographic compositions of the test takers are similar before and after the introduction of the RP option. Examinees under the two testing modes differ slightly on certain background variables. Once the group differences are adjusted by propensity score weighting, the estimated mode effects are small and nonsystematic across test titles overall. We note some variations across subgroups based on gender and race.

Keywords At-home testing; remote proctoring; propensity score weighting; test taker demographics; performance patterns; test taker participation; mode effects; sensitivity analysis

doi:10.1002/ets2.12386

Introduction

At-home testing via remote proctoring (RP) was introduced in spring 2020 by many testing organizations during the widespread shutdown caused by the COVID-19 pandemic (Camara, 2020); this dual-option practice will probably continue. Although examinees appreciate the convenience of test taking without leaving their homes, the coexistence of two test modes creates a scenario for potential fairness issues (Hu, 2020). The choice of RP testing may be affected by factors unrelated to the test construct, as RP testing has technical requirements for equipment and internet connection, as well as requiring access to a quiet personal space for an extended period of time (Muckle et al., 2022). Test takers may also choose to take tests at home, which causes a lower level of anxiety (Stowell & Bennett, 2010). The reverse could also be true, as faculty and students initially raised concerns over student privacy and added test anxiety brought about by remote monitoring (Flaherty, 2020; Patil & Bromwich, 2020). The two self-selected groups are probably not equivalent, and any observed difference in the test outcome is a compound of many factors (Brown et al., 2023). Careful investigation is needed to separate the self-selected group differences from the observed outcome differences to detect the test mode effect (ME) or other factors of interest.

Studies of test score comparability have been conducted at different stages in the history of testing. In the days of paper-based testing (PBT), there were studies of the equivalence of alternate test forms, such as test forms with item scrambling. Special data collection designs (e.g., equivalent groups) and equating were used to adjust for possible differences to ensure equivalence of test scores (Harris, 1991). When computer-based testing (CBT) was first introduced, studies were conducted addressing the comparability of test scores from CBT and PBT (Jeong, 2014). Studies were also conducted to address the comparability of test taking on different devices (Chen et al., 2014) or with different presentation features, such as screen size and screen resolution (Bridgeman et al., 2001). The common goals in this line of inquiry are to establish evidence that test scores carry the same meaning across different testing situations and to detect and control for factors contributing to any differential effects related to test conditions (Bennett, 2003).

In the ideal scenario, pilot studies are conducted to randomly assign test takers to alternate testing formats or conditions so that direct comparison can be made of test outcomes across conditions. In most practical situations, however, random

Corresponding author: J. Miao, E-mail: jmiao@ncsbn.org

assignment is not an option. For example, in K – 12 state assessments, school districts may administer tests with either PBT or CBT using difference devices, mostly based on their resources and everyday instructional practices. In such cases, statistical methods are available to remove the impact of non-content-related factors. Duque (2016)'s research employed scores from Partnership for Assessment of Readiness for College and Careers (PARCC). Duque used hierarchical linear modeling to compare differences in PARCC scores based on test mode, with students nested in schools controlling for student characteristics such as race/ethnicity, gender, special education, learner status, and prior achievement. Liu et al. (2016) studied PARCC mode effects by matching schools within each state based on student demographic characteristics using propensity score matching.

Not all changes in testing were planned, and the COVID-19 pandemic presented a "salient example of unexpected disruption" (Baldwin & Clauser, 2022, p. 140). The expansion of at-home testing via remote proctoring sparked another round of interest in studying score comparability. Puhan and Kim (2022) summarized statistical procedures that could be used to evaluate potential mode effects at both the item level and the total test level. They also compared the linking relationship between the test center (TC) and the at-home testing groups to determine the reporting score conversion from a subpopulation invariance perspective. Jones et al. (2022) used two methods to detect mode effects. First, they calibrated the same pool of items separately using two TC cohorts and one online proctoring (OP) cohort and found the results to be more similar between the two TC cohorts, therefore indicating mode effects. For the second method, the authors studied repeat test takers in either the same or a different modality, forming two pairs of repeaters (TC – TC vs. TC – OP and OP – OP vs. OP – TC), matched on their first-attempt scores. All four groups had increased pass rates, and a noticeable mode effect was found in both pairs, but less strongly in the OP – TC condition.

Kim and Walker (2021) used a pseudo-equivalent groups (PEG) approach to examine TC and RP comparability at both the item score and test score levels for three licensure tests. They used examinee background information to construct weights via minimum discriminant information adjustment (MDIA; see Haberman, 2014, 2015) to transform self-selected groups into pseudo-equivalent groups. At the item level, they compared TC versus RP item difficulty differences from two methods (i.e., delta adjustment and PEG adjustment) to detect mode effect. Their results show small nonsystematic differences between TC and RP. At the test level, the PEG-adjusted RP group conversion was compared to the original TC group conversion; the differences were found to be small, leading to the same pass/fail outcome for most test takers. The study found small and nonsignificant mode effects for the test titles investigated.

These earlier studies were conducted shortly after the COVID-19 pandemic and focused on the total test taker group. The current study used data over a longer time period (from September 2016 to August 2022) to identify the pattern of test participation and performance under the TC and RP dual-mode condition. Test participation pattern was defined as the proportion of examinees choosing the RP option. Test performance was measured by mean scale score and pass rate. We used propensity score weighting (PSW) based on existing background information to balance the TC and RP groups before estimating the mode effect. We also conducted sensitivity analyses to assess the potential impact of omitted variables (OVs) on the results. We studied the pattern in the total group as well as in subgroups based on gender and race to address four questions:

1. Is there any significant difference in the demographic composition of the test-taking population before and after the introduction of RP testing?
2. What is the participation and performance pattern of TC versus RP in the test-taking population?
3. Is there evidence for a systematic and substantial test mode effect?
4. Is the TC versus RP test performance pattern consistent across subgroups based on gender and race?

Data

We conducted preliminary analyses on data from 10 licensure tests and decided to include 3 in this study based on several considerations. First, these tests have large enough test taker volumes to allow for in-depth analysis at the subgroup level. Second, the tests are of distinctive content areas and therefore have somewhat different test taker populations. The three tests also have different test formats. Tests 1 and 3 include multiple-choice (MC) items only, and Test 2 includes both MC items and constructed-response items. Table 1 lists the main features of the three selected test titles.

The RP test option first started in May 2020, toward the end of the 2019 – 2020 testing year; however, we included data for six testing years, from September 2016 to August 2022, to provide a broader context for understanding the impact of

Table 1 Features of the Three Studied Tests

	Test 1	Test 2	Test 3
Test format	MC only	MC + CR	MC only
Average annual volume	4,700	15,000	3,400
First timers (%)	76	90	58
Scale score range	100 – 200	100 – 200	100 – 200
Median score	170	176	165
Standard error of measurement	5.5	5.1	5.6
Standard error of scoring	NA	2.1	NA
Launch of remote proctoring test option	May 2020	June 2020	September 2020

Note. CR = constructed response. MC = multiple choice. NA = not applicable.

Table 2 Demographic Composition of First-Time Test Takers From September 2016 to August 2022

	Test 1		Test 2		Test 3	
	Before RP launch	Since RP launch	Before RP launch	Since RP launch	Before RP launch	Since RP launch
Total group	11,836	9,474	43,985	23,618	7,872	3,779
Gender (%)						
Male	4	3	10	11	59	59
Female	96	97	90	89	41	41
Race/ethnicity (%)						
African American	15	16	7	8	14	16
Hispanic	8	7	3	5	3	3
White	64	66	80	77	72	70

Note. RP = remote proctoring.

the change. We included only first-time candidates to remove the confounding factors of repeated attempts (e.g., practice effect). We then focused on the dual test option period for TC and RP comparisons, propensity score weighting, and mode effect estimation.

Table 2 provides sample sizes and the demographic composition[1] for the studied period. The results indicate that gender and race composition is similar before and after the introduction of RP. The three studied tests are in different content areas, and their test candidate populations are somewhat different. The Test 1 candidate pool is overwhelmingly female, whereas the Test 3 candidate pool includes more men than women. Both Test 1 and Test 3 have higher percentages of African American test takers than Test 2, and Test 1 also has a higher percentage of Hispanic test takers.

Methods

For each test, we first examined descriptive statistics to identify the overall pattern of participation and mean performance before and after the launch of RP testing. We then focused on the period with both test options to examine the patterns for TC and RP modes. We calculated the mean scaled score and pass rate[2] by test mode (TC or RP) for the total group as well as by subgroup based on gender and race. We also calculated the RP participation rate in each group.

Since examinees self-selected to take TC or RP, the observed performance differences between TC and RP are confounded with group differences. Kim and Walker (2021) used the MDIA approach to adjust for group differences and create pseudo-equivalent groups. We used propensity score weighting in this study, which is more accessible with multiple statistical packages available (Keller & Tipton, 2016; SAS Institute, 2016). The propensity score, $p(X_i)$, is defined by Rosenbaum and Rubin (1983) as the probability of group membership, r_i, given a set of covariates or background variables, X_i. Propensity scores can be estimated using logistic regression (PSMATCH in SAS) or other methods, such as the generalized boosted models in the R twang package (McCaffrey et al., 2004; Ridgeway, 2006; Ridgeway et al., 2023).

The covariates in this study were self-reported information that candidates provided during test registration, including date of birth, gender, race/ethnicity, linguistic background, educational background (e.g., undergraduate major and grade point average), types of teacher training programs (e.g., master's program or fifth-year program), years of teaching

experience, teaching career plan, geographical region (urban, suburban, or rural), and so on. These responses were dummy coded for the propensity score analysis. Appendix A provides a list of all covariates considered in this study.

We chose propensity score *weighting* instead of propensity score *matching* so that we could keep all eligible cases in the analysis.[3] We used the R twang package to estimate the average treatment effect in the treated (ATT) weights and matched the RP group to the TC group; that is, we took the TC as the baseline condition and compared the weighted RP results to the observed TC results to assess the mode effect. In this scenario, the TC group is the target group ($r_i = 1$) and the RP group is the control group ($r_i = 0$). Each TC case has a weight of 1, and each RP case has a weight of $w(x_i)$, calculated using

$$w\left(x_i\right) = \frac{p\left(x_i\right)}{1 - p\left(x_i\right)}. \tag{1}$$

We examined distributions of all covariates for TC and RP groups and found the distributions to be very similar. We focused on the mean to summarize the differences and reported the pre- and post- weighting standardized mean difference (SMD) between TC and RP groups for each covariate in the model to evaluate group balance. The SMD is estimated by

$$\text{SMD} = \frac{\hat{\mu}_{\text{TC}} - \hat{\mu}_{\text{RP}}}{\hat{\sigma}} \tag{2}$$

We then estimated the mode effect (ME) in the outcome measures (y_i), that is, mean test score and pass rate, using

$$\text{ME} = \frac{\sum_{i=1}^{N} r_i y_i}{\sum_{i=1}^{N} r_i} - \frac{\sum_{i=1}^{N} \left(1 - r_i\right) w\left(x_i\right) y_i}{\sum_{i=1}^{N} \left(1 - r_i\right) w\left(x_i\right)}. \tag{3}$$

Because r_i is either 0 or 1, the first term in Equation 3 is the average outcome of the TC group. The second term is the weighted average of the RP group, as the $(1 - r_i)$ term selects out all the TC group cases.

We estimated the mode effect using the mean scale score and pass rate in the total group as well as in subgroups by gender (male and female) and race (White, African American, and Hispanic). A positive ME indicates that the TC group has higher performance, whereas a negative ME means that the RP group has higher performance.

Ideally, the estimated ME is close to 0; if that is not the case, then either there is a substantial mode effect, or some important differences are not captured by the weights. The effectiveness of propensity score weighting depends on the availability of relevant and adequate background variables. We used the R OVtool package (Burgette et al., 2022; Griffi et al., 2020) to assess the sensitivity of the mode effect estimates and statistical significance to unobserved variables based on characteristics of observed ones. Basically, the tool illustrates how the estimated effect would change if an omitted variable were included in the propensity score model. The assumption is that the omitted variable is similar to an observed variable in terms of its relationship with the group membership and the outcome.

Results

We present the results for each test, providing (a) descriptive statistics on test participation and performance, (b) SMD on covariates before and after weighting to assess group balance, (c) estimated mode effect with PSW, and (d) sensitivity to omitted variables. We then summarize the total group results across the three tests. Additionally, we explore alternative models for the African American group on Test 3 to address the subgroup differences.

Test 1 Total Group Results

Figure 1 shows 6-year trends in test volume and mean scale score for Test 1. Test volume had been declining for 4 consecutive years, with a bigger drop in 2019–2020, probably due to the COVID-19 pandemic. It bounced back by almost 50% in 2020–2021, then returned to the 2016–2017 level in 2021–2022. The mean scale score is stable across the 6 years, with a slight increase in the first 3 years and a slight decrease in the last 2 years. Since the launch of RP in May 2020, approximately 61% of first-time test takers have chosen the RP option.

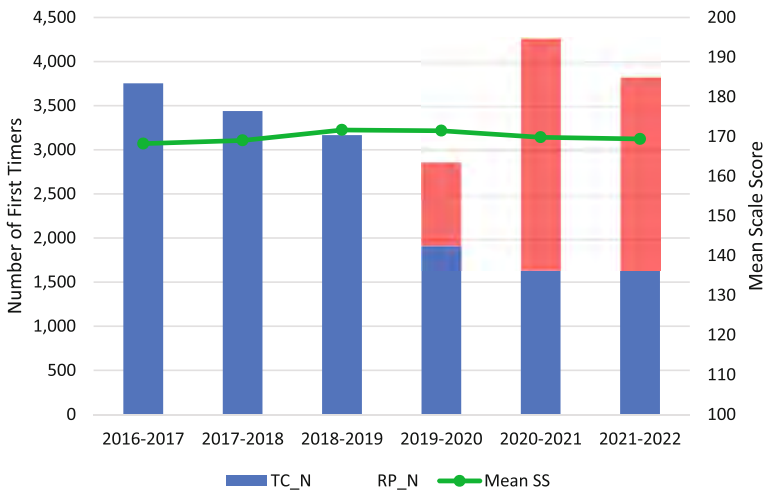

Figure 1 Test 1 first-time participation and mean performance by testing year. RP = remote proctoring. SS = scale score. TC = test center.

Table 3 Observed Test Outcome by Test Center and Remote Proctoring: Test 1 (TC – RP)

			Scale score				Pass rate		
Group	*N*	RP%	*M*	*SD*	Mean difference	Effect size	Pass rate (%)	Difference (%)	Effect size
Total					−2.01**	−0.13		−3.63**	−0.11
TC	3,722		168.6	16.1			84.6		
RP	5,752	61	170.6	14.6			88.2		
Male					−1.85	−0.12		−2.46	−0.07
TC	137		167.3	17.5			81.8		
RP	190	58	169.2	15.7			84.2		
Female					−2.01**	−0.13		−3.65**	−0.11
TC	3,585		168.7	16.1			84.7		
RP	5,562	61	170.7	14.5			88.4		
White					−0.46	−0.04		−0.82	−0.03
TC	2,329		173.2	12.7			93.3		
RP	3,948	63	173.7	12.1			94.2		
African American					−1.88**	−0.11		−3.37	−0.07
TC	662		158.3	17.4			64.1		
RP	841	56	160.2	17.1			67.4		
Hispanic					−3.39**	−0.20		−2.79	−0.06
TC	298		160.2	18.1			72.8		
RP	369	55	163.6	15.5			75.6		

Note. RP = remote proctoring. TC = test center. *Unweighted difference statistically significant at $p < .05$. **Unweighted difference statistically significant at $p < .01$.

Observed Test Outcome by TC and RP

Table 3 provides the observed test outcomes on Test 1 by test mode (TC or RP) for the total group as well as for the subgroups based on gender and race since the launch of RP in May 2020. The RP participation rate is 61% in the total group, with small variations across subgroups, that is, male (58%) lower than female (61%) and African American (56%) and Hispanic (55%) lower than White (63%). Overall, the RP group has a slightly higher average performance than the TC group in terms of mean scale score (2.01 points) and pass rate (3.63%). Women account for 96% of the first-timer test takers on Test 1, with a slightly larger observed TC – RP difference than observed for men. The White group has a much smaller observed TC – RP performance difference than the African American group and the Hispanic group. The observed differences are statistically significant for the total group and some subgroups, but the effect sizes are small (⩽.20).

Figure 2 shows the scale score distributions by TC and RP for the total group as well as for subgroups based on gender and race. For Test 1, the RP group test performance is slightly higher than it is for the TC group. The observed TC and RP scale score distributions appear to be very close for the total group, with some variations across the subgroups.

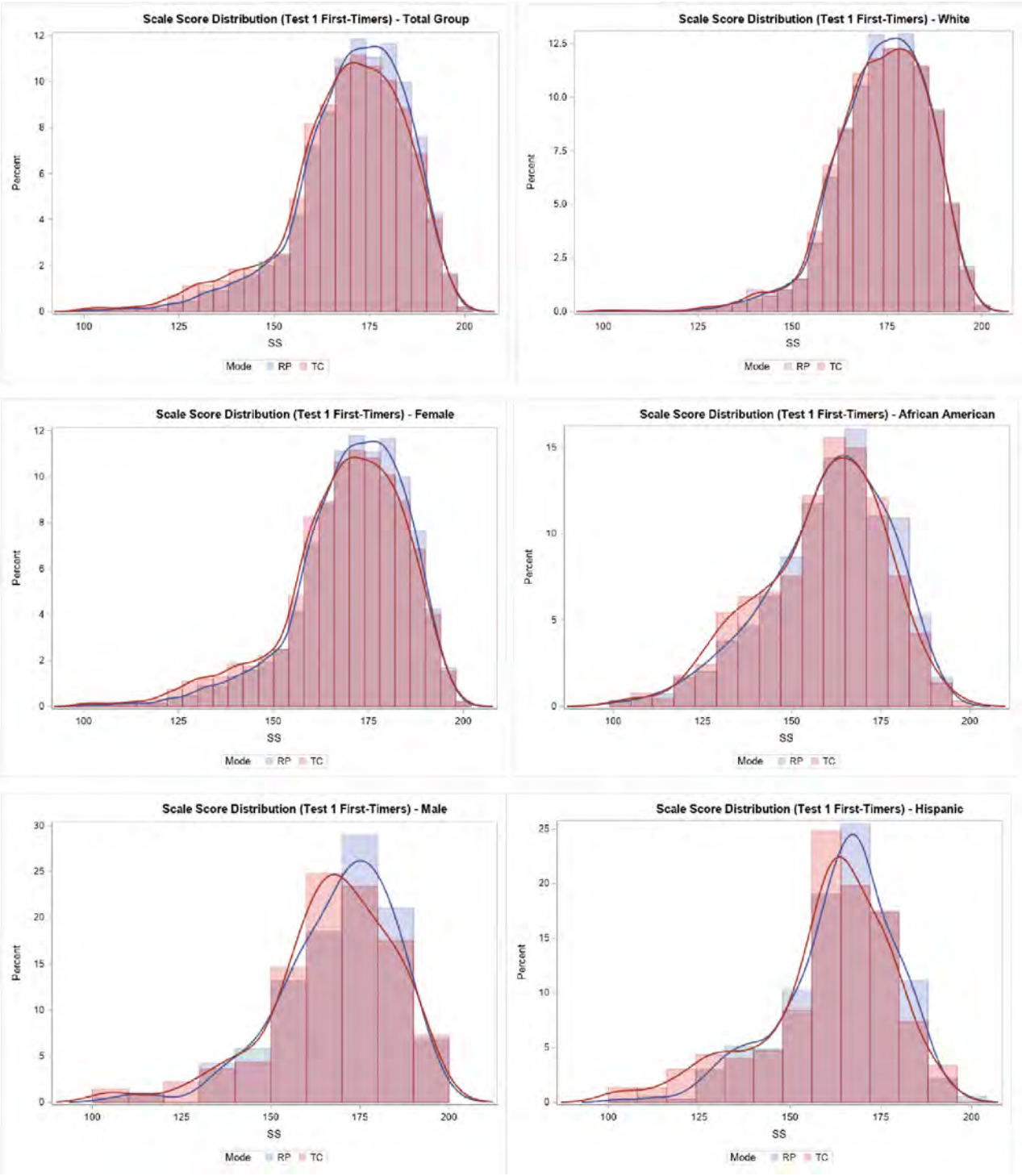

Figure 2 Test 1 scale score distributions by test mode. RP = remote proctoring. TC = test center.

Group Balance

Figure 3 displays the SMDs between the RP and TC groups on the covariates included in the propensity score model for Test 1. Positive SMDs indicate larger values for the TC group. In general, effect sizes under .20 are considered small. The unweighted SMDs are <.10 in absolute value for all except two variables. Relatively larger group differences are observed

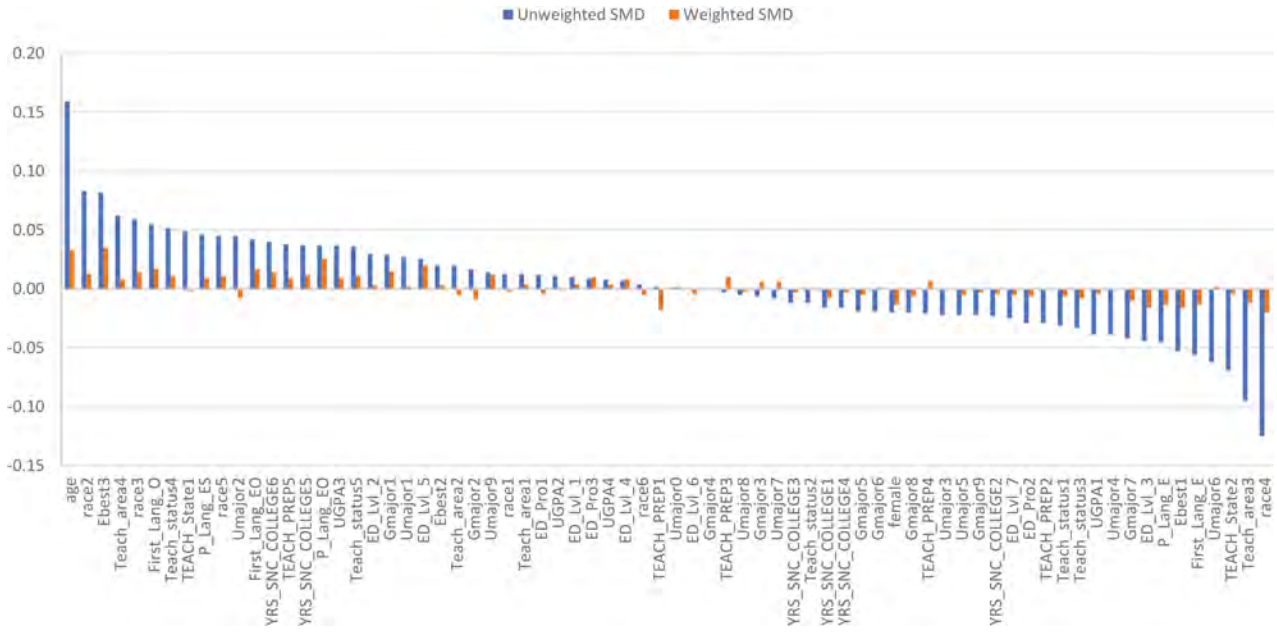

Figure 3 Standardized mean differences on covariates for Test 1. SMD = standardized mean difference.

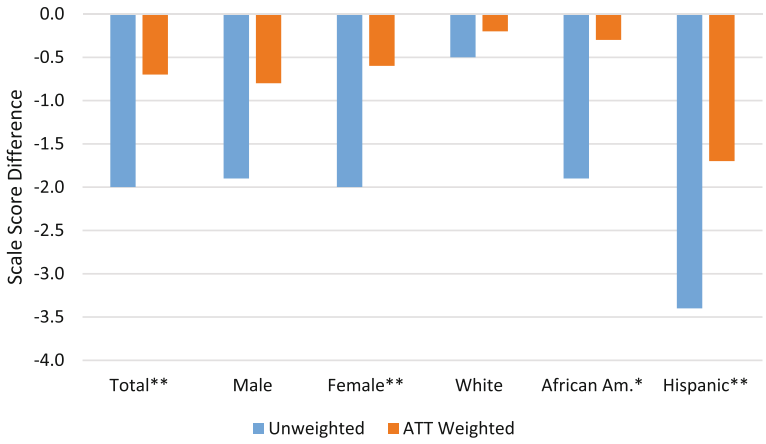

Figure 4 Test 1 mean scale score difference (TC − RP). ATT = average treatment effect in the treated. *Unweighted difference statistically significant at $p < .05$. **Unweighted difference statistically significant at $p < .01$.

for two variables, indicating that the RP group is younger (i.e., age) and has a higher percentage of test candidates who self-identified as White (i.e., race4). The weighted SMDs are reduced to close to zero, indicating a good balance between the two groups after weighting. Table B1 in Appendix B provides both the weighted and unweighted SMDs for all covariates in the propensity score model for Test 1.

Weighted Test Outcome by TC and RP

With the PSW results, we can estimate the mode effect for Test 1. Figure 4 displays the mean scale score differences. For context, the scale score standard deviation is approximately 15 for Test 1. Overall, RP performance is higher than TC performance. For the total group, the observed mean score difference is approximately 2 points. The observed mean difference is similar across gender, smaller for the White group, and larger for the Hispanic group. The observed score difference is significant, $p < .05$, for all groups except male and White. The ATT weighted mean score difference is smaller for each group and not statistically significant.

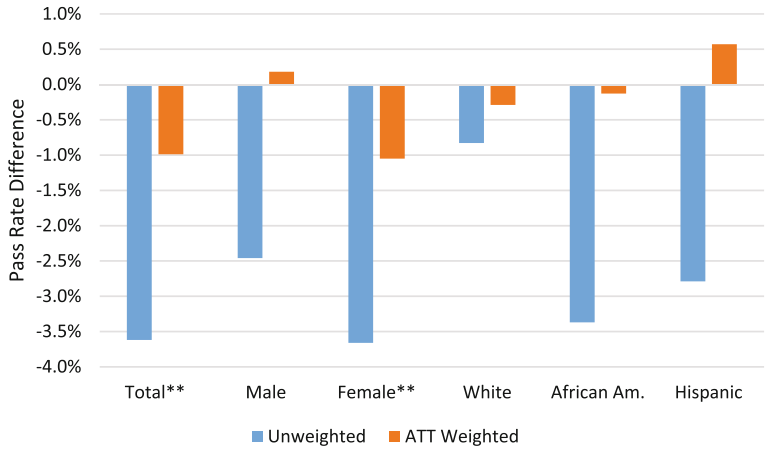

Figure 5 Test 1 pass rate difference (TC − RP). ATT = average treatment effect in the treated. *Unweighted difference statistically significant at $p < .05$. **Unweighted difference statistically significant at $p < .01$.

Figure 5 displays the pass rate differences. Overall, RP performance is higher than TC performance. For the total group, the observed pass rate difference is approximately 3.6%. The observed pass rate difference is significant, $p < .05$, for the total group and the female group. The ATT weighted pass rate difference is smaller for each group and not statistically significant. This pattern is consistent across subgroups, with some small variations. Table B2 in Appendix B provides the weighted differences in mean scale score and pass rate in the total group as well as in subgroups based on gender and race for Test 1.

Sensitivity Analyses

Figure 6 illustrates the sensitivity analyses for Test 1, where the estimated mode effect on scale score is −0.564 (i.e., TC performance is slightly lower than RP performance), $p = .121$. The figure shows how the estimated effect (indicated by the black solid contours) and the p-value (indicated by the red contours) would change as a function of an OV that is somewhat similar to the observed covariates. The X-axis indicates the association with treatment (i.e., choice of TC or RP) expressed as the SMD of the OV between the TC and RP groups, and the Y-axis is the correlation between the OV and the outcome (i.e., scale score). The black solid contours represent the adjusted effect estimates. The red contours show how statistical significance is impacted. The blue dots represent OVs similar to the observed variables. For example, if an OV similar to race2 were added to the model, the estimated effect would be close to −1.0 with $p \approx .01$. On the other hand, if an OV similar to race4 were added, it could change the estimated effect to roughly above 0.1 with $p > .10$. In Figure 6, most blue dots concentrate in the lower center area of the plot, indicating a weaker relationship to the outcome and with little impact on the estimated ME. Based on Figure 6, we can conclude that the Test 1 results are reasonably robust to OVs and that the estimated ME is trustworthy; that is, TC performance is slightly lower (by <1.0 point) than RP performance for Test 1.

Test 2 Total Group Results

Figure 7 shows the 6-year trends in test volume and mean scale score for Test 2, which are very similar to the trends for Test 1. The annual test volume declined for the first 4 years, with a bigger drop in 2019–2020 due to the COVID-19 pandemic. It bounced back by an increase of over 36% in 2020–2021, then fell slightly in 2021–2022. The mean scale score is stable, with slight decreases in 2020–2021 and 2021–2022. Since the launch of RP in June 2020, approximately 57% of first-time test takers have chosen the RP option.

Observed Test Outcome by TC and RP

Table 4 provides the observed test outcome on Test 2 by test mode (TC or RP) for the total group as well as for the subgroups based on gender and race since the launch of RP in June 2020. The RP participation rate is 57% in the total

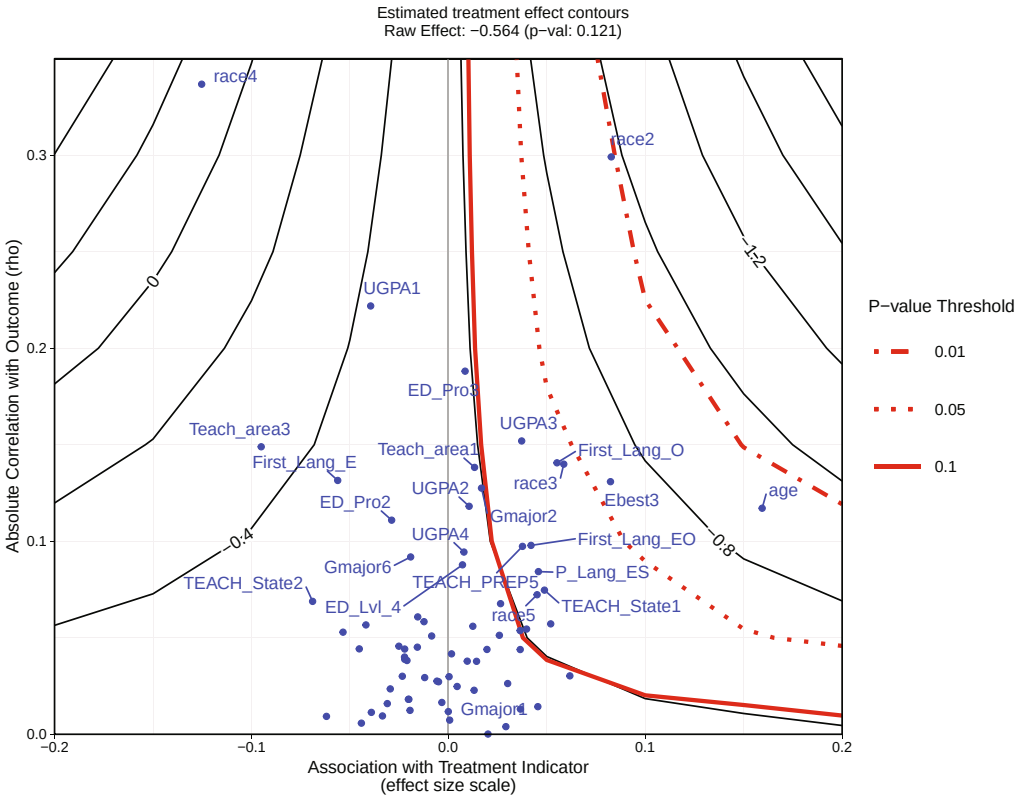

Figure 6 OVtool sensitivity plot for Test 1.

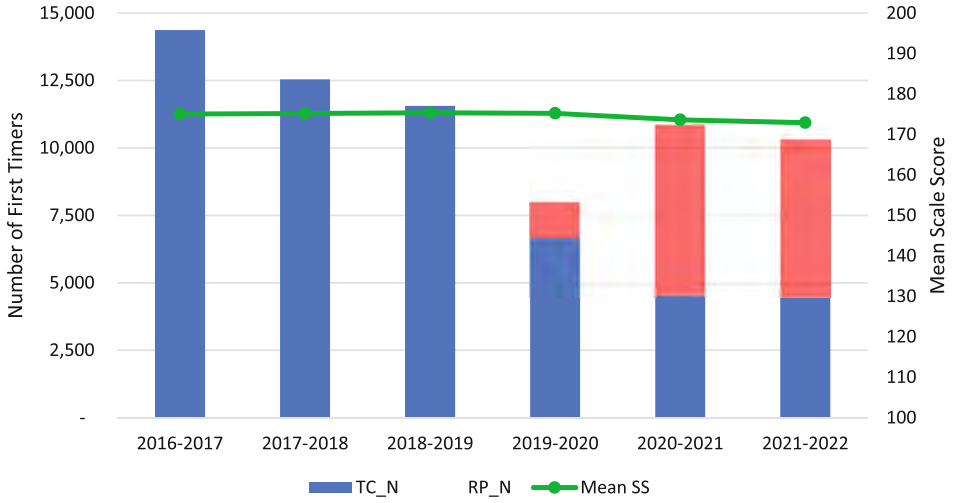

Figure 7 Test 2 first-time participation and mean performance by testing year. RP = remote proctoring. SS = scale score. TC = test center.

group, with very small variations across subgroups: male (55%) lower than female (57%), and African American (58%) slightly higher than White and Hispanic (both 57%). Overall, the TC group performance is slightly higher than the RP group performance in terms of mean scale score (0.40 points) and pass rate (0.47%). For female first-time test takers, the TC group has slightly higher performance, whereas in the male group, RP performance is slightly higher. In the White group, TC performance is slightly higher, whereas for the African American and Hispanic groups, RP performance is slightly higher. The observed differences are statistically significant for the total group and some subgroups, but the effect sizes are small (≤.10).

Table 4 Observed Test Outcome by Test Center and Remote Proctoring: Test 2 (TC − RP)

			Scale score				Pass rate		
Group	*N*	RP%	*M*	*SD*	Mean difference	Effect size	Pass rate (%)	Difference (%)	Effect size
Total					0.40**	0.04			
TC	10,199		173.8	11.1			88.6	0.47	0.01
RP	13,419	57	173.4	11.4			88.1		
Male					−0.16	−0.02			
TC	1,171		171.0	12.0			82.6	−1.10	−0.03
RP	1,452	55	171.2	12.3			83.7		
Female					0.49**	0.04			
TC	9,028		174.1	11.0			89.4	0.71	0.02
RP	11,967	57	173.6	11.3			88.7		
White					0.76**	0.07			
TC	7,855		174.9	10.4			91.5	1.62*	0.06
RP	10,422	57	174.2	10.8			89.9		
African American					−1.11	−0.09			
TC	765		167.4	12.2			73.5	−4.26*	−0.10
RP	1,055	58	168.5	13.4			77.7		
Hispanic					−0.45	−0.04			
TC	485		168.7	12.9			77.5	−1.41	−0.03
RP	641	57	169.2	13.4			78.9		

Note. RP = remote proctoring. TC = test center. *Unweighted difference statistically significant at $p < .05$. **Unweighted difference statistically significant at $p < .01$.

Figure 8 shows the scale score distributions by TC and RP for the total group as well as for subgroups based on gender and race. For Test 2, the observed test performance distributions in the total group are almost identical for TC and RP; however, there are greater variations in the smaller subgroups (i.e., male, African American, Hispanic).

Group Balance

Figure 9 displays the SMDs between the RP and TC groups on the covariates included in the propensity score model for Test 2. The unweighted SMDs are mostly <.10 in absolute value, except for a few variables. Compared to the TC group, the RP group has a lower percentage of candidates who plan to enroll or were enrolled in a teacher preparation program (i.e., teach_status1), and they are less likely to get teacher preparation through an undergraduate degree (i.e., teach_prep1). The RP group has more recent college graduates during the last 1 – 3 years (i.e., YRS_SNC_COLLEGE2) and a higher percentage with 1 – 3 years of teaching experience (i.e., teach_status3). Table B3 in Appendix B provides both the weighted and unweighted SMDs for all covariates in the propensity score model for Test 2.

Weighted Test Outcome by TC and RP

We can now use the PSW results to estimate the mode effect for Test 2. Figure 10 displays the mean scale score differences. For context, the scale score standard deviation is approximately 11 for Test 2. For the total group, TC performance is slightly higher than RP performance. The observed score difference is less than 1 scale score point for all groups, except African Americans. The magnitude of the ATT weighted differences is slightly larger for some groups but still all within 1 point. There is some variation across subgroups; in particular, RP group performance is higher than TC group performance for the African American and Hispanic groups. The differences are statistically significant, $p < .01$, for the total group, the female group (unweighted only), and the White group, probably owing to the larger sample sizes.

Figure 11 displays the pass rate differences. The pattern is very similar to the scale score differences. Table B4 in Appendix B provides the weighted differences in mean scale score and pass rate in the total group as well as in subgroups based on gender and race for Test 2.

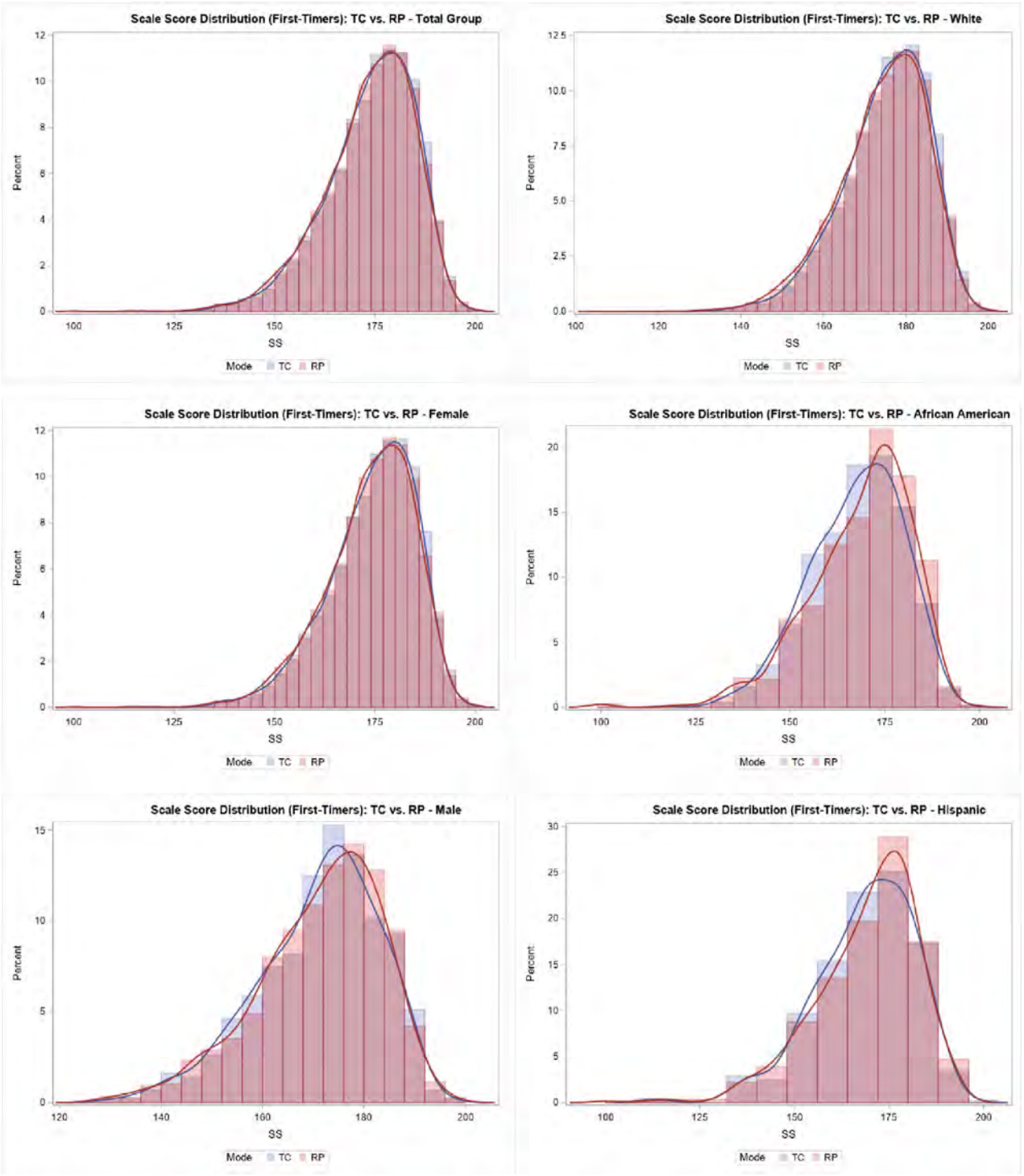

Figure 8 Test 2 scale score distributions by test mode. RP = remote proctoring. TC = test center.

Sensitivity Analyses

Figure 12 illustrates the sensitivity analyses for Test 2, where the estimated mode effect on scale score is 0.569 (i.e., TC performance slightly higher than RP performance) and is statistically significant. The blue dots, which represent OVs similar to the observed variables, concentrate in the lower right side of the plot, indicating that the OVs could reduce the estimated effect somewhat. For example, if an OV similar to UGPA1 were added, the estimated effect could change

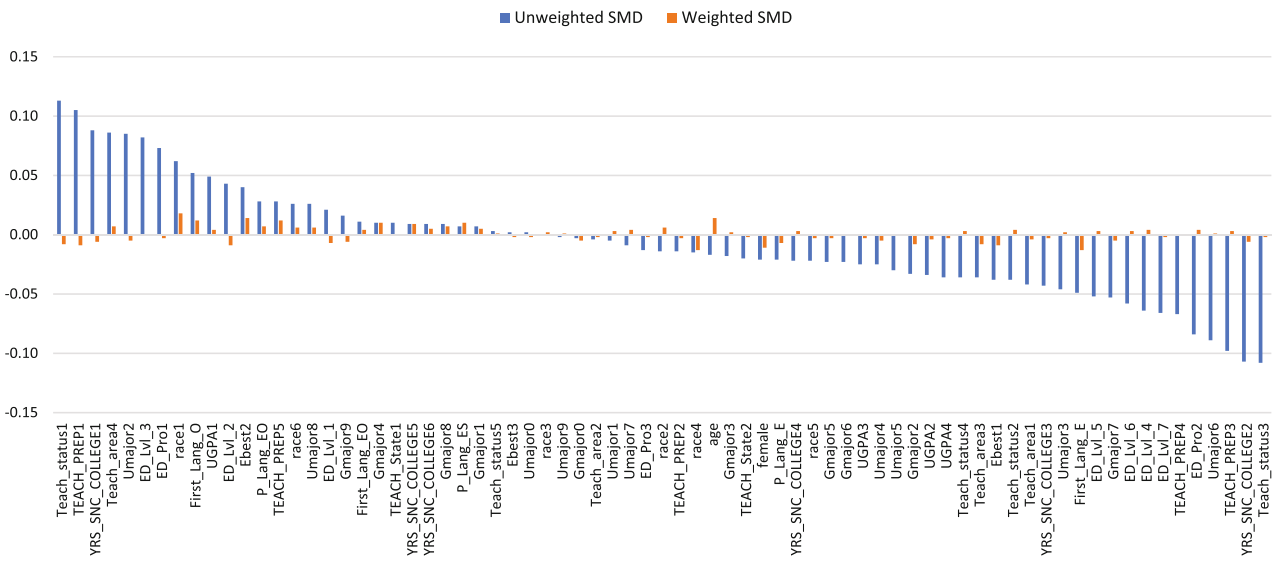

Figure 9 Standardized mean differences on covariates for Test 2. SMD = standardized mean difference.

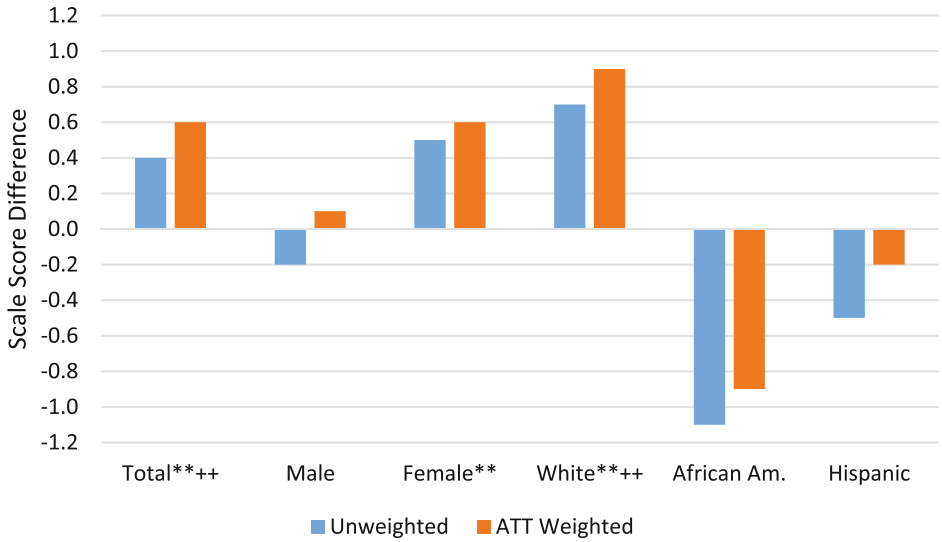

Figure 10 Test 2 mean scale score difference (TC − RP). ATT = average treatment effect in the treated. **Unweighted difference statistically significant at $p < .01$. ++Weighted difference statistically significant at $p < .01$.

to roughly 0.45 scale score point with $p < .01$. Based on Figure 12, we can conclude that the Test 2 results are reasonably robust to OVs and that the estimated ME is trustworthy; that is, TC performance is slightly higher (by <0.6 point) than RP performance for Test 2.

Test 3 Total Group Results

Figure 13 shows the 6-year trends in test volume and mean scale score for Test 3. The annual test volume was relatively stable across the first 3 years, reaching a peak in 2018–2019. Test volume showed a noticeable dip of almost 30% in 2019–2020 due to the COVID-19 pandemic. Test volume bounced back in 2020–2021, then fell again slightly in 2021–2022. The mean scale score has a very slight downward trend until 2019–2020, then rises slightly over the next 2 years.

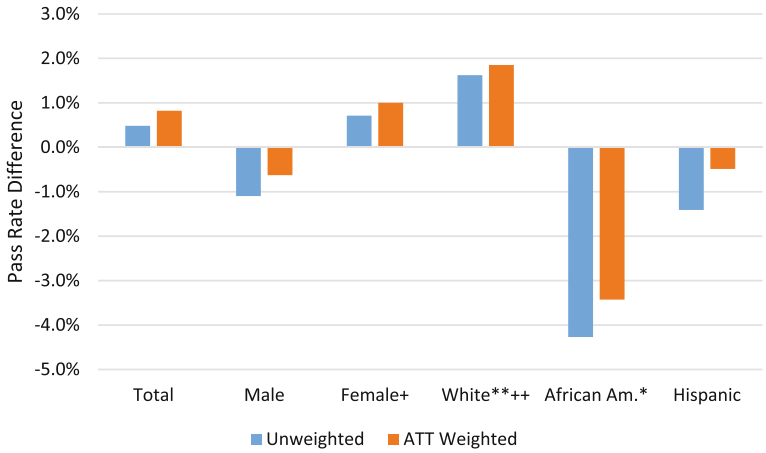

Figure 11 Test 2 pass rate difference (TC − RP). ATT = average treatment effect in the treated. *Unweighted difference statistically significant at $p < .05$. **Unweighted difference statistically significant at $p < .01$. $^{+}$Weighted difference statistically significant at $p < .05$. $^{++}$Weighted difference statistically significant at $p < .01$.

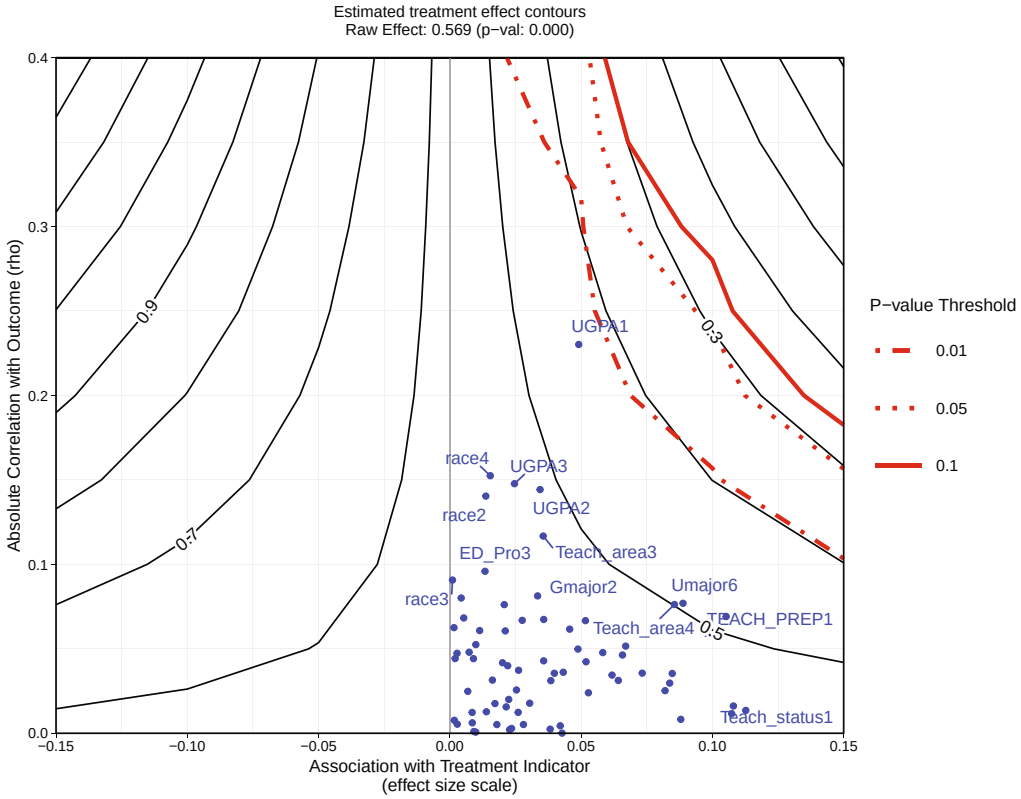

Figure 12 OVtool sensitivity plot for Test 2.

Observed Test Outcome by TC and RP

Table 5 provides the observed test outcome on Test 3 by test mode (TC or RP) for the total group as well as for the subgroups based on gender and race since the launch of RP in September 2020. The RP participation rate is 61% in the total group, with small variations across subgroups: male (60%) lower than female (62%), and Hispanic (60%) slightly lower than White and African American (both 61%). Overall, the RP group has slightly higher average performance than the TC group in terms of mean scale score (1.05 score point) and pass rate (1.75%). The female group has a slightly smaller observed TC − RP difference than the male group. The White group has a much smaller observed performance difference

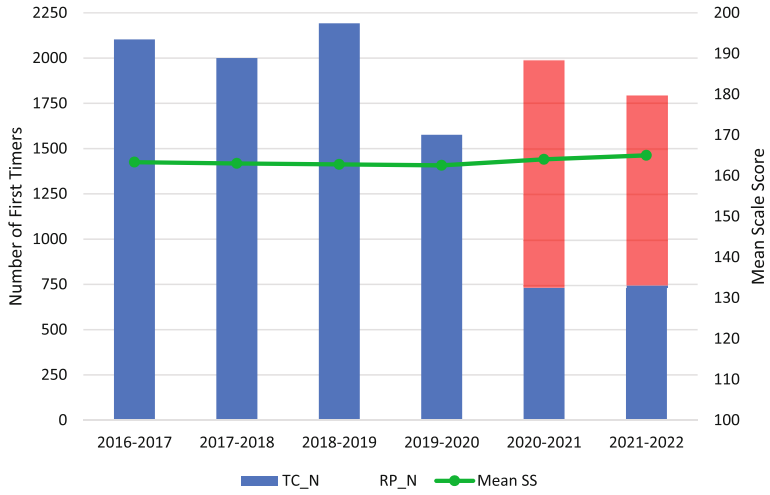

Figure 13 Test 3 first-time participation and mean performance by testing year. RP = remote proctoring. SS = scale score. TC = test center.

Table 5 Observed Test Outcome by Test Center and Remote Proctoring: Test 3 (TC – RP)

			Scale score				Pass rate		
Group	*N*	RP%	*M*	*SD*	Mean difference	Effect size	Pass rate (%)	Difference (%)	Effect size
Total					−1.05*	−0.08			
TC	1,488		163.9	13.4			71.7	−1.75	−0.04
RP	2,291	61	164.9	12.5			73.5		
Male					−1.19*	−0.09			
TC	889		162.9	13.6			69.0	−2.53	−0.06
RP	1,322	60	164.1	13.0			71.5		
Female					−0.75	−0.07			
TC	599		165.3	13.0			75.8	−0.37	−0.01
RP	969	62	166.1	11.7			76.2		
White					−0.52	−0.05			
TC	1,031		166.5	11.4			80.1	0.02	−0.00
RP	1,603	61	167.1	10.6			80.1		
African American					−4.37**	−0.28			
TC	236		152.1	15.2			36.9	−13.27**	−0.27
RP	363	61	156.5	15.7			50.1		
Hispanic					−2.05	−0.17			
TC	44		160.8	13.7			61.4	−2.27	−0.05
RP	66	60	162.9	12.0			63.6		

Note. RP = remote proctoring. TC = test center. *Unweighted difference statistically significant at $p < .05$. **Unweighted difference statistically significant at $p < .01$.

than the African American group and the Hispanic group. The observed differences are statistically significant in some groups, with mostly small effect sizes, except for the African American group.

Figure 14 shows the scale score distributions by TC and RP for the total group as well as for subgroups based on gender and race. For Test 3, RP performance is slightly higher than TC performance in general; however, the TC and RP score distributions differ noticeably from each other for the African American group. *One may ask whether this noticeable difference is due to self-selection bias or test mode effect*.

Group Balance

Figure 15 displays the SMDs between the RP and TC groups on the covariates included in the propensity score model for Test 3. The unweighted SMDs are all <.10 in absolute value, except for one variable (i.e., Umajor9), indicating that the RP group is more likely than the TC group to have an undecided undergraduate major. Table B5 in Appendix B provides both the weighted and unweighted SMDs for all covariates in the propensity score model for Test 3.

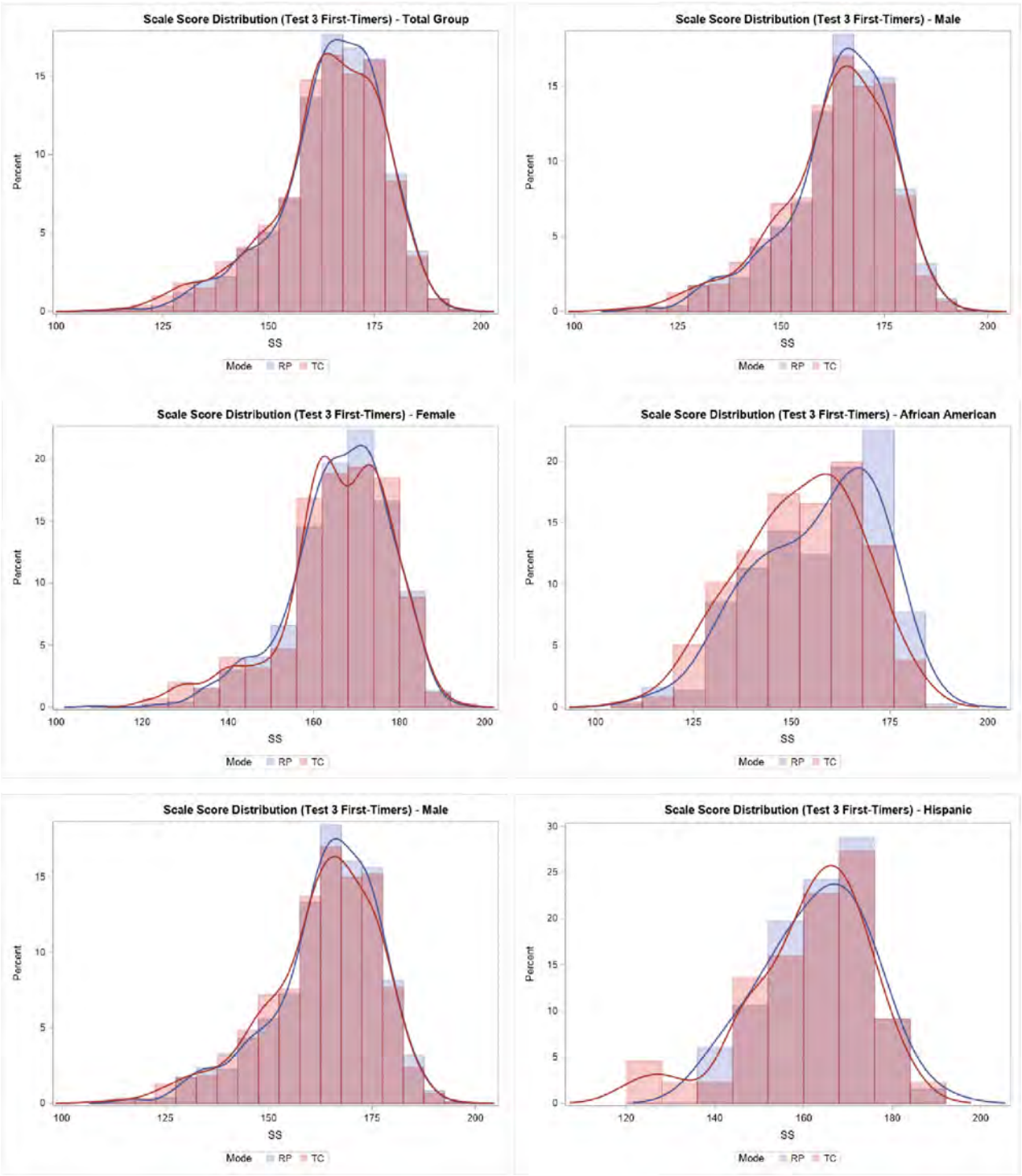

Figure 14 Test 3 scale score distributions by test mode. RP = remote proctoring. TC = test center.

Weighted Test Outcome by TC and RP

We can now use the PSW results to estimate the mode effect for Test 3. Figure 16 displays the mean scale score differences. For context, the scale score standard deviation is approximately 13 for Test 3. RP performance is higher than TC performance for all groups. The observed mean score difference is approximately 1 point at the total group level; the weighted

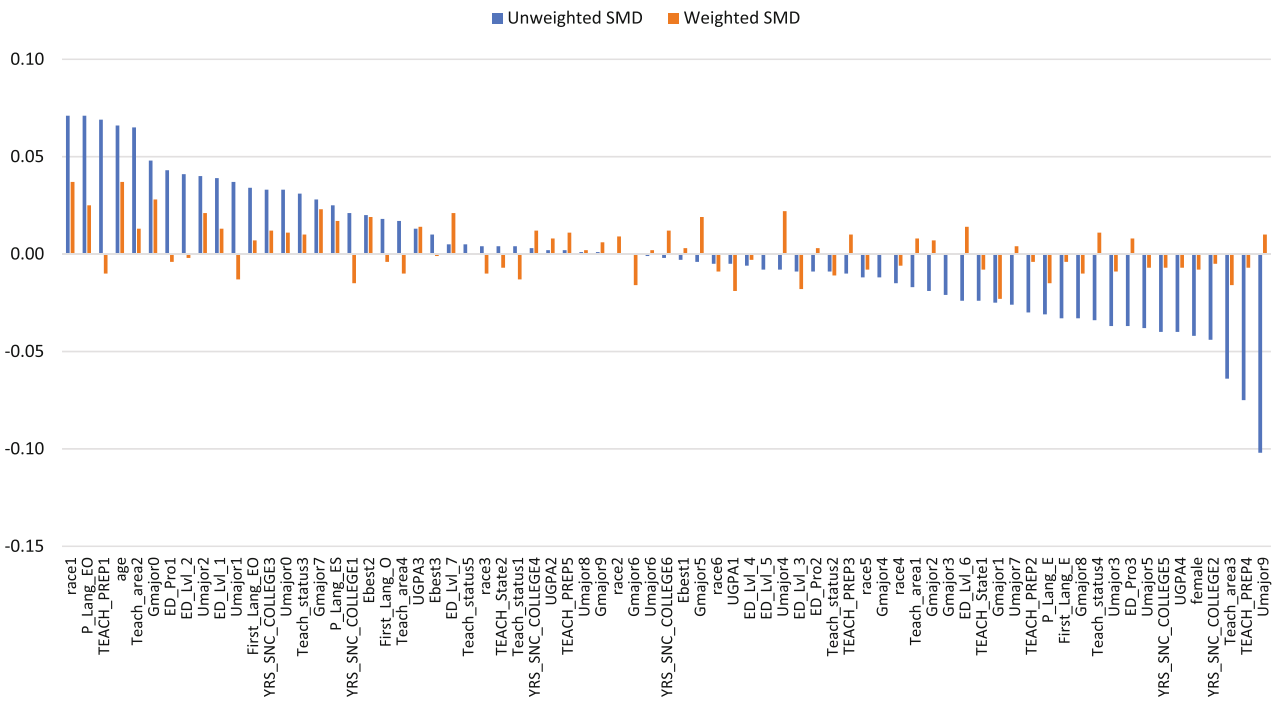

Figure 15 Standardized mean differences on covariates for Test 3, total group model. SMD = standardized mean difference.

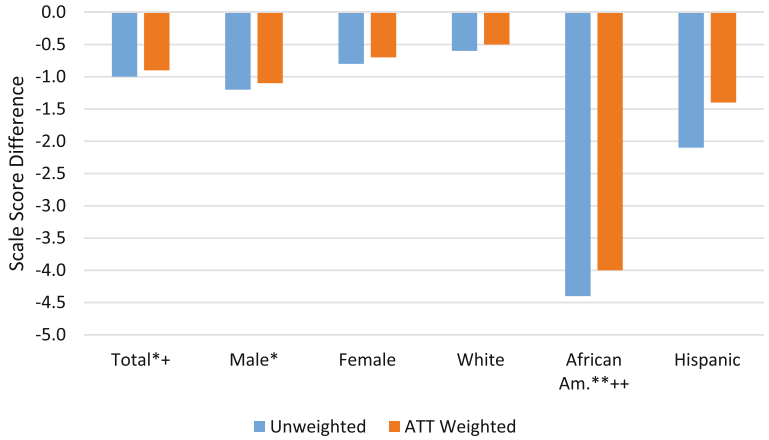

Figure 16 Test 3 mean scale score difference (TC − RP). ATT = average treatment effect in the treated. *Unweighted difference statistically significant at $p < .05$. **Unweighted difference statistically significant at $p < .01$. $^{+}$Weighted difference statistically significant at $p < .05$. $^{++}$Weighted difference statistically significant at $p < .01$.

difference is smaller, but not by much. This pattern is consistent across subgroups; however, the performance difference is much larger for the African American group.

Figure 17 displays the pass rate differences, with a pattern similar to the scale score differences. The pass rate difference for the African American group is much larger than the difference in other subgroups. Table B6 in Appendix B provides the TC − RP weighted differences in mean scale score and pass rate in the total group as well as in subgroups based on gender and race for Test 3.

Sensitivity Analyses

Figure 18 illustrates the sensitivity analyses for Test 3, where the estimated mode effect on scale score is −0.931 (i.e., TC performance is lower than RP performance), $p = .038$. The blue dots represent OVs similar to the observed variables,

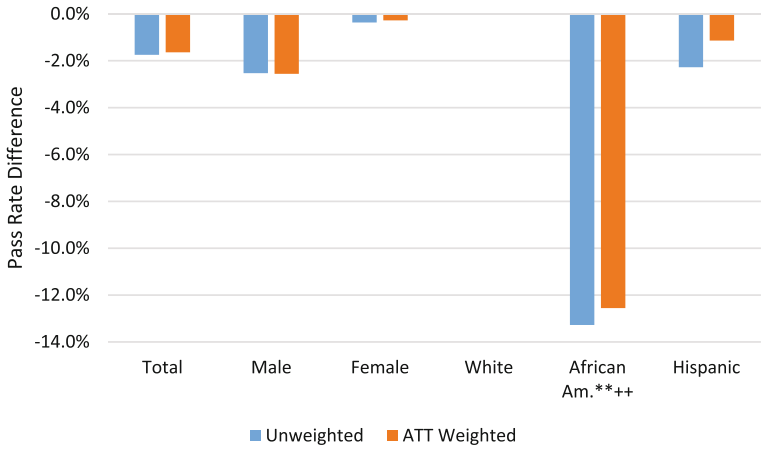

Figure 17 Test 3 pass rate difference (TC − RP). ATT = average treatment effect in the treated. **Unweighted difference statistically significant at $p < .01$. $^{++}$Weighted difference statistically significant at $p < .01$.

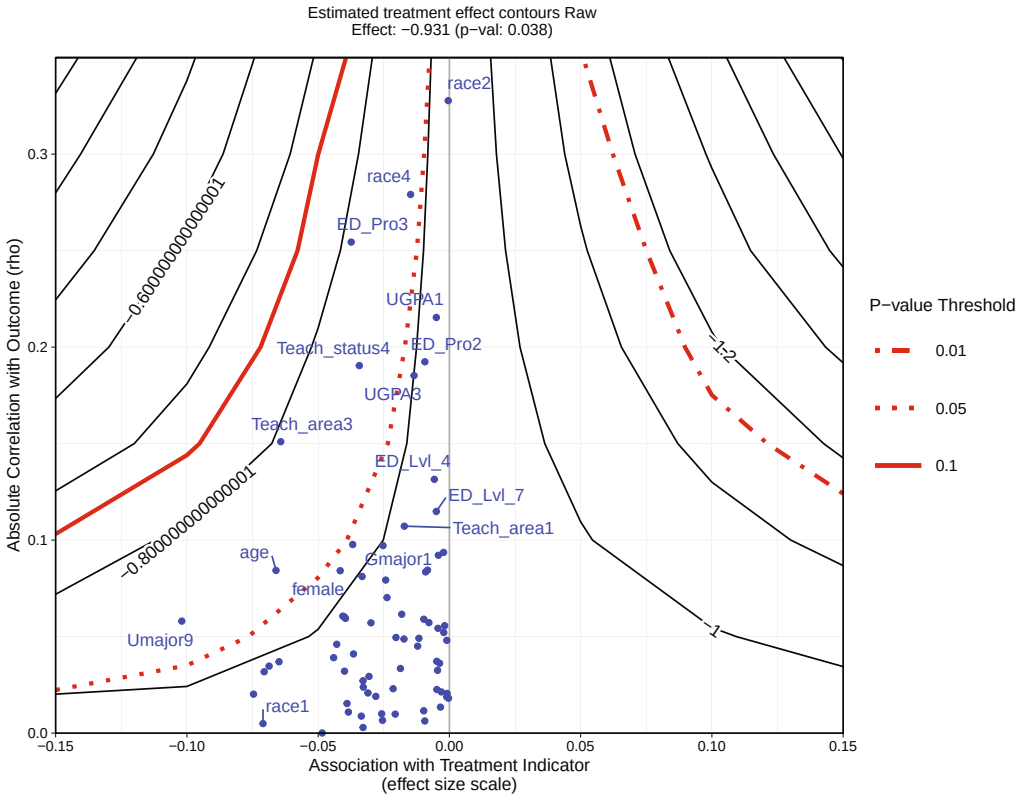

Figure 18 OVtool sensitivity plot for Test 3, total group.

and they concentrate in the center, around the −0.9 black contour line. If an OV similar to Teach_area3 were added, the estimated effect could change to roughly −0.80 scale score point with $p < .10$. Based on Figure 18, we can conclude that the Test 3 results are reasonably robust to OVs and that the estimated ME is trustworthy at the total group level; that is, TC performance is slightly lower (by <1 point) than RP performance for Test 3. However, we could not help but wonder about the much bigger estimated effect in the African American group from the preceding analyses.

Summary of Total Group Propensity Score Weighting Results

Across the three tests included in this study, we can make a few general observations on the total group PSW model results:

- The demographic composition of the examinee population, in terms of gender and race, is similar before and after the launch of the RP test option.
- From the beginning of RP testing (launched in May, June, and September 2020) until August 2022, more than half of first-time examinees have chosen RP over TC. The RP participation rate is fairly consistent across gender and race groups.
- The observed performance differences between the TC and RP groups are generally small and inconsistent across tests. The performance difference is consistent across gender groups but tends to be larger for the non-White groups, which have much smaller sample sizes.
- The TC and RP groups have similar distributions on the background variables with very small SMDs (i.e., mostly <.10). The PSW effectively adjusts the group differences on existing covariates, creating pseudo equivalent groups (in terms of the background variables).
- At the total group level, the estimated mode effects on mean scale score are less than 1 scale score point (−0.6, 0.6, and − 0.9, respectively) for the three studied tests, and the estimated mode effects for pass rates are less than 2% (−1.0%, 0.8%, and − 1.7%, respectively) for the three tests. These results provide evidence for small and inconsistent mode effects.
- At the subgroup level, the estimated mode effects show some variation.

Test 3 African American Subgroup Results

The large estimated effect in the African American group for Test 3 is intriguing. One wonders if the bigger difference is due to differential mode effect or if some important variables have been missing from the PSW model. We explored the existing data more closely and found that African American candidates for Test 3 appear to concentrate in a small number of states, whereas similar concentrations are not observed in the total group. We included additional variables in alternative propensity score models representing these states: attending institution (AI) and designated institution (DI; the institution where test scores would be sent) for Arkansas (AR), Louisiana (LA), North Carolina (NC), and Virginia (VA); see Table B7 in Appendix B for a complete list of variables for Test 3 African American subgroup analysis.

Alternative Propensity Score Weighting Models

We compared estimated effects for African American group using three alternative models: Model 1 and Model 2 are both based on the total group data, but Model 2 includes additional state covariates. Model 3 also includes additional state covariates but is based on the African American group data only. Table 6 provides the estimated effects using weights from these three models. Of the two total group–based models, the estimated ME from Model 2 is somewhat smaller with the additional state variables. The subgroup-based Model 3 yields the smallest estimated mode effect.

Group Balance

Figure 19 displays the SMDs between the RP and TC groups on the covariates included in Model 3. We see several effect sizes above .10, including several state variables (DI_VA, AI_VA, DI_LA, and AI_LA). The weighted SMDs are closer to zero, indicating improved balance between the two groups. Table B7 provides the SMDs for all covariates in Test 3, Model 3. It is noted that the African American group is much smaller ($n = 599$) than the total group ($N = 3{,}779$) and therefore more prone to random error.

Table 6 Estimated Effects From Alternative Models for Test 3, African American

			SS		Pass rate	
Model	Sample	Covariates	Difference: weighted	*p*-Value	Difference: weighted (%)	*p*-Value
1	Total	Without states	−4.02	0.003	−12.65	0.003
2	Total	With states	−3.38	0.012	−9.81	0.022
3	Afr. Am.	With states	−2.86	0.063	−9.74	0.041

Note. SS = scale score.

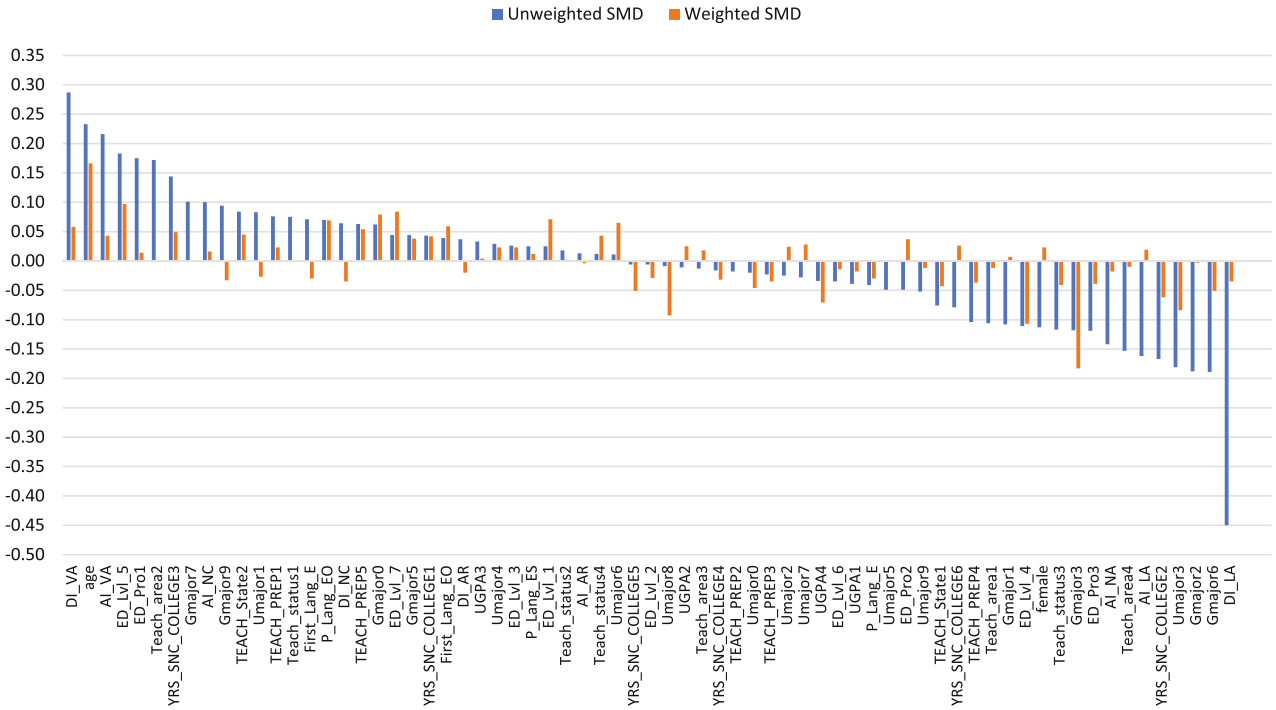

Figure 19 Standardized mean differences on covariates: Test 3, African American subgroup model. SMD = standardized mean difference.

Sensitivity Analyses

Figure 20 illustrates the sensitivity analyses for Test 3, Model 3, where the estimated ME on scale score is −2.86 (i.e., TC performance lower than RP performance). The blue dots representing the OVs concentrate mostly in the center, between the black contour lines of −3.2 and − 2.6. One state variable (DI_LA) is on the far left side, meaning that if an OV similar to DI_LA were added to the model, the estimated effect would change to below −2.4 with $p > .10$. Based on Figure 20, we can conclude that the Model 3 results are reasonably robust to OVs that are similar to the observed covariates. TC performance is lower (by about 3 points) than RP performance for the African American group on Test 3.

Discussion

As a response to COVID-19, many testing programs introduced RP testing in 2020, and this has become regular practice. Examinees self-selected to take the TC or RP option instead of being randomly assigned, so the two groups are likely to be nonequivalent in their backgrounds (demographic, academic, or professional), ability levels, and test performance. To ensure test validity, scores from different testing conditions must be comparable, without being affected by test mode effects or other related concerns, such as test security breaches.

Perhaps the most effective way to test for mode effects is by randomizing testing conditions across a large group of test takers and comparing results across modes. Such a study is rarely feasible, and when it is, it can be quite costly. A large motivation of this study was to find a method of simulating a randomized trial that used existing data, was widely familiar to researchers, was relatively easy to implement, and for which software was readily available. We have not seen another study with all of these characteristics. For that reason, we chose propensity score weighting, which can be fitted through many statistical packages.

Another motivation was to examine possible effects within subgroups. We believed that although MEs may not manifest at the total group level, there could still be evidence of MEs in subgroups of interest. Such evidence could raise a fairness issue, meriting further examination.

In this study, we used propensity score weights derived from background variables to balance the groups and then estimated the test mode effect. We also assessed the sensitivity of the estimated effect to the impact of omitted variables. We studied the mode effect in the total group as well as in subgroups based on gender and race. We found small and

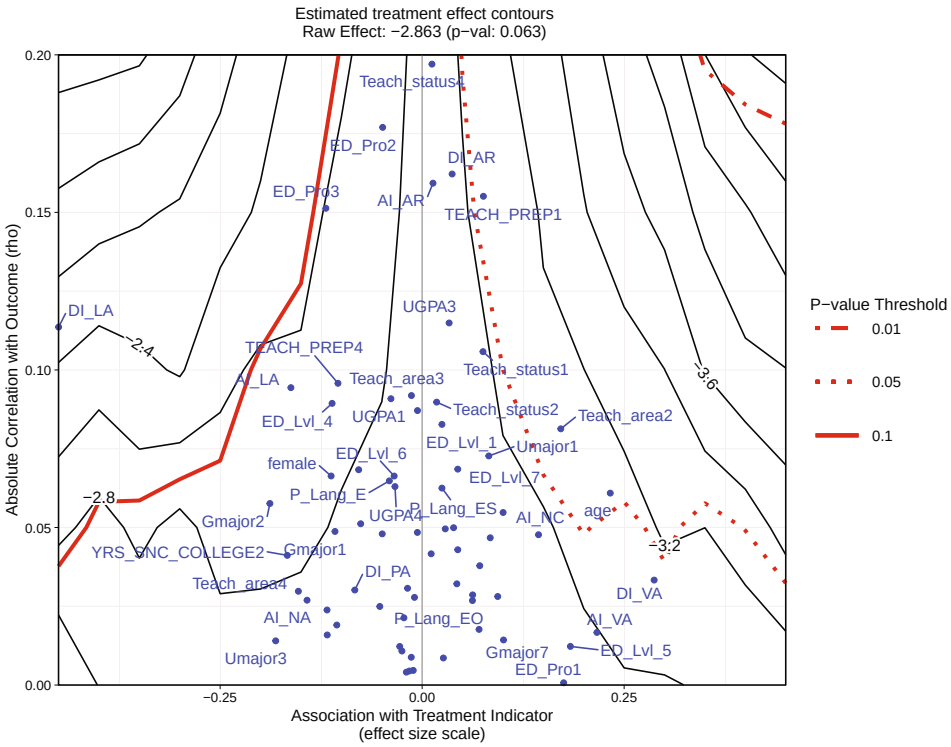

Figure 20 OVtool sensitivity plot for Test 3, African American subgroup.

inconsistent effects in the total group across the three tests. Within subgroups, we found more variation. One observation that stands out at the subgroup level is of the larger performance differences across modes for non-White groups, such as for the African American subgroup taking Test 3. In this subgroup, RP performance was on average 4 scale score points higher than TC performance (see Figure 14). We found that by including additional state variables and using subgroup-only data for the PSW model, the estimated effect was lowered to 2.86. This result suggests that differences in examinee characteristics might be contributing to observed mode differences, rather than the modes themselves. On the other hand, the sensitivity analyses suggest that the results are robust, and the pattern is unlikely to be changed by an OV.

One important condition for propensity score modeling is the "no unmeasured confounders" assumption, in that all variables that affect treatment assignment and outcome have been measured (Austin, 2011). The covariates in this study were collected during test registration from responses to 16 background questions, including questions about age, gender, race, linguistic background, educational background, teacher training experience, and geographic area for teaching. Propensity score weighting is most effective when the covariates are strongly related to the outcome of interest. We conducted three sets of regression analyses to check for the strength of the relationship between covariates to the choice of test mode and test outcome. First, the ordinary least squares regression of the test score on all covariates in the total group yielded R^2 values ranging from .173 to .317 (adjusted $R^2 = .168–.304$) for the three tests (see Table B8). Second, we conducted logistic regression of the pass/fail outcome on the covariates, and the percentage correctly classified ranged from 73% to 83% (see Table B9). We also conducted logistic regression of TC/RP group membership on the covariates, yielding 57% to 59% of examinees correctly classified (see Table B10). Given this information, we could conclude that the PSW approach partially adjusted for group ability differences; the estimated effect was "cleaner" but not "pure." Although the sensitivity analyses indicate that the results are reasonably robust, we were only able to assess OVs *similar to* the observed ones.

The regression (of test outcome on covariates) results in Tables B8 and B9 show that the models fit slightly better in the TC setting than in the RP setting. This is reasonable, as the background questions were originally developed for the TC context. Given the unplanned nature of the introduction of RP testing during the pandemic, many factors that may

have affected the choice of the RP option were not included, such as adequate hardware and software, a reliable internet connection, a suitable home environment, and comfort with visual monitoring via the computer camera.

The logistic regression (of test mode choice on covariates) models (Table B10) included different numbers of covariates for each test, indicating that tests may have their own unique sets of covariates and that one uniform background survey may not capture the unique factors relevant to each content area. Examination of standardized mean differences on covariates (see Figures 3, 9, and 15) indicates that the TC and RP groups differ on different characteristics for each of the three tests (i.e., age and race2 for Test 1, teaching status and teacher preparation for Test 2, and undergraduate major for Test 3). For Test 3, state turned out to be an important factor for the African American test takers but not for the total group. More customized background data collection would be helpful but challenging to implement in operational settings.

In this study, we found small mode effects in either direction, meaning that the effect is context specific, without any clear advantage for either condition. Critics of RP testing are quick to question the security of at-home testing; although there is always a possibility of security breaches, we found no evidence of increased test anomalies in RP testing based on our operational monitoring. Figures B1 – B3 in Appendix B show the mean scale scores of first-time test takers by testing year, and the subgroup mean scores show a consistent pattern with the total group over time. This suggests that the dual-mode option does not have a differential effect across subgroups.

It is plausible that the larger estimated effect in the subgroup could be caused by the larger disparity within groups on the missing covariates that affect both test mode choice and test outcome. For example, socioeconomic status (SES) has long been documented as positively related to academic performance. SES also directly affects access to RP testing, as examinees need the right equipment, a quiet personal space, and a reliable internet connection. A larger SES disparity between the TC and RP groups for non-White groups could contribute to the larger effect observed in this study.

Another unaccounted-for factor could be test anxiety, which is found to be related to cultural context, gender, and age (Lowe, 2019; Torrano et al., 2020) and has a negative impact on academic performance (Chang, 2021; Torrano et al., 2020; Woldeab & Brothen, 2019). Spence et al. (2019) found that despite the moderate number of problems with RP, 55% – 73% of participants indicated that RP would reduce their anxiety on future examinations. Given the choice of their preferred test mode, there may be different levels of test anxiety between TC and RP groups that may affect test performance but are not captured in the background information.

In this study, we were able to balance the TC and RP groups on existing covariates to create pseudo equivalent groups to evaluate test mode effects. Overall, the estimated effects were small and nonsystematic. However, the covariates most likely only partially adjusted for group ability differences. Some important covariates (e.g., SES, test anxiety level) were not available, and these unmeasured factors may affect both test mode choice and test performance. Such information would help to parse out the differences in test outcomes but is hard to collect operationally. The sensitivity analyses provided a method to frame the impact of omitted variables, and results in this study are fairly robust, suggesting the potential influence to be somewhat limited. One could study repeat test takers across different test modes if the group sizes are sufficient for PSW or other methods. Future studies of test taker experience would also shed light on the dual test mode practice and help testing organizations improve access and ensure comparability and validity of reported test scores. As the memory of the COVID-19 pandemic fades from the public consciousness, motivations for choosing TC or RP testing may continue to evolve. Thus periodic evaluation of the mode effect is recommended as long as both options are offered.

Acknowledgment

The authors thank Dan McCaffrey for his expertise and assistance in the methods and analyses of this study. Jing Miao is currently a senior psychometrician at the National Council of State Boards of Nursing (NCSBN), Chicago, Illinois, United States.

Endnotes

1 Groups with fewer than 100 examinees (from the RP launch until August 2022) are not included.
2 Each licensing state decides on its own cut score, although a recommended score is provided based on multistate standard setting. In this study, we apply the most used (often the same as the recommended) cut score to calculate pass rate.
3 Additionally, in attempting to approximate random assignment, propensity score matching often leads to greater rather than lesser imbalance (King & Nielsen, 2019).

References

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*(3), 399–424. https://doi.org/10.1080/00273171.2011.568786

Baldwin, P., & Clauser, B. E. (2022). Historical perspectives on score comparability issues raised by innovations in testing. *Journal of Educational Measurement*, *59*(2), 140–160. https://doi.org/10.1111/jedm.12318

Bennett, R. E. (2003). *Online assessment and the comparability of score meaning* (Research Memorandum No. RM-03-05). ETS. https://www.ets.org/Media/Research/pdf/RM-03-05-Bennett.pdf

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2001). *Effects of screen size, screen resolution, and display rate on computer-based test performance* (Research Report No. RR-01-23). ETS. https://doi.org/10.1002/j.2333-8504.2001.tb01865.x

Brown, M. I., Grossenbacher, A., & Warman, Z. (2023). Self-selection as an explanation for general mental ability test score differences between mobile and nonmobile devices in observational studies. *Journal of Applied Psychology*, *108*(7), 1190–1206. https://doi.org/10.1037/apl0001067

Burgette, L., Pane, J., Griffin, B. A., & McCaffrey, D. (2022, October 12). *Package "OVtool."* https://cran.r-project.org/web/packages/OVtool/OVtool.pdf

Camara, W. (2020). Never let a crisis go to waste: Large-scale assessment and the response to COVID-19. *Educational Measurement: Issues and Practice*, *39*(3), 10–18. https://doi.org/10.1111/emip.12358

Chang, Y. F. (2021). 2-dimentional cognitive test anxieties and their relationships with achievement goals, cognitive resources, motivational engagement and academic performance. *Learning and Individual Difference*, *92*, Article 102084. https://doi.org/10.1016/j.lindif.2021.102084

Chen, G., Cheng, W., Chang, T.-W., Zheng, X., & Huang, R. (2014). A comparison of reading comprehension across the paper, computer screens, and tablets: Does tablet familiarity matter? *Journal of Computers in Education*, *1*, 213–225. https://doi.org/10.1007/s40692-014-0012-z

Duque, M. (2016). *Is there a PARCC mode effect?* (SDP Fellowship Capstone Report 2016). https://sdp.cepr.harvard.edu/files/cepr-sdp/files/sdp-fellowship-capstone-parcc-mode.pdf

Flaherty, C. (2020, May 10). Big proctor. *Inside Higher Ed*. https://www.insidehighered.com/news/2020/05/11/online-proctoring-surging-during-covid-19

Griffin, B. A., Ayer, L., Pane, J., Vegetabile, B., Burgette, L., McCaffrey, D., Coffman, D. L., Cefalu, M., Funk, R., & Godley, M. D. (2020). Expanding outcomes when considering the relative effectiveness of two evidence-based outpatient treatment programs for adolescents. *Journal of Substance Abuse Treatment*, *118*, Article 108075. https://doi.org/10.1016/j.jsat.2020.108075

Haberman, S. J. (2014). *A program for adjustment by minimum discriminant information* (Research Memorandum No. RM-14-01). ETS. https://www.ets.org/research/policy_research_reports/publications/report/2014/jscd.html

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

Harris, D. (1991). Effects of passage and item scrambling on equating relationships. *Applied Psychological Measurement*, *15*(3), 247–256. https://doi.org/10.1177/014662169101500304

Hu, J. C. (2020, June 24). Graduate programs drop GRE after online version raises concerns about fairness. *Science*. https://www.sciencemag.org/careers/2020/06/graduate-programs-drop-gre-after-online-version-raises-concerns-about-fairness

Jeong, H. (2014). A comparative study of scores on computer-based tests and paper-based tests. *Behavior and Information Technology*, *33*(4), 410–422. https://doi.org/10.1080/0144929X.2012.710647

Jones, P., Tong, Y., Liu, J., Borglum, J., & Primoli, V. (2022). Score comparability between online proctored and in-person credentialing exams. *Journal of Educational Measurement*, *59*(2), 180–207. https://doi.org/10.1111/jedm.12320

Keller, B., & Tipton, E. (2016). Propensity score analysis in R: A software review. *Journal of Educational and Behavioral Statistics*, *41*(3), 326–348. https://doi.org/10.3102/1076998616631744

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/ets.2.12323

King, G., & Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political Analysis*, *27*(4), 435–454. https://doi.org/10.1017/pan.2019.11

Liu, J., Brown, T., Chen, J., Ali, U., Hu, L., & Costanzo, K. (2016). *PARCC Mode Comparability Study based on spring 2015 operational test data* (ED599049). ERIC. https://files.eric.ed.gov/fulltext/ED599049.pdf

Lowe, P. A. (2019). Exploring cross-cultural and gender differences in test anxiety among U.S. and Canadian college students. *Journal of Psychoeducational Assessment*, *37*(1), 112–118. https://doi.org/10.1177/0734282917724904

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, *9*(4), 403–425. https://doi.org/10.1037/1082-989X.9.4.403

Muckle, T. J., Meng, Y., & Johnson, S. (2022). A quantitative evaluation of a live remote proctoring pilot. *Journal of Applied Testing Technology*, *23*, 46–53. https://jattjournal.net/index.php/atp/article/view/165806/115524

Patil, A., & Bromwich, J. E. (2020, September 29). How it feels when software watches you take tests. *New York Times*. https://www.nytimes.com/2020/09/29/style/testing-schools-proctorio.html

Puhan, G., & Kim, S. (2022). Score comparability issues with at-home testing and how to address them. *Journal of Educational Measurement*, *59*(2), 161–179. https://doi.org/10.1111/jedm.12324

Ridgeway, G. (2006). Assessing the effect of race bias in post–traffic stop outcomes using propensity scores. *Journal of Quantitative Criminology*, *22*(1), 1–29. https://doi.org/10.1007/s10940-005-9000-9

Ridgeway, G., McCaffrey, D., Morral, A., Cefalu, M., Burgette, L., Pane, J., & Griffin, B. A. (2023). *Toolkit for weighting and analysis of nonequivalent groups: A guide to the twang package*. https://cran.r-project.org/web/packages/twang/vignettes/twang.pdf

Rosenbaum, P., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

SAS Institute. (2016). *SAS/STAT® 14.2 user's guide: The PSMATCH procedure*. https://support.sas.com/documentation/onlinedoc/stat/142/psmatch.pdf

Spence, D., Ward, R., Wooden, S., Browne, M., Song, H., Hawkins, R., & Wojnakowski, M. (2019). Use of resources and method of proctoring during the NBCRNA continued professional certification assessment: Analysis of outcomes. *Journal of Nursing Regulation*, *10*(3), 37–46. https://doi.org/10.1016/S2155-8256(19)30147-4

Stowell, J., & Bennett, D. (2010). Effects of online testing on student exam performance and test anxiety. *Journal of Educational Computing Research*, *42*(2), 161–171. https://doi.org/10.2190/EC.42.2.b

Torrano, R., Ortigosa, J. M., Riquelme, A., Mendez, F. J., & Lopez-Pina, J. A. (2020). Test anxiety in adolescent students: Different response according to the components of anxiety as a function of sociodemographic and academic variables. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.612270

Woldeab, D., & Brothen, T. (2019). Online proctoring, test anxiety, and student performance. *International Journal of E-Learning and Distance Education*, *34*(1). ERIC. https://files.eric.ed.gov/fulltext/EJ1227595.pdf

Appendix A

List of All Covariates

Variable	Description
age	Calculated based on date of birth
female	Gender = F
AI State	State of the attending institute (i.e., where the candidate goes to school)
AI_AR	AI = Arkansas
AI_LA	AI = Louisiana
AI_NA	AI = not available
AI_NC	AI = North Carolina
AI_VA	AI = Virginia
DI State	State of the designated institute (i.e., where the candidate seeks license)
DI_AR	DI = Arkansas
DI_LA	DI = Louisiana
DI_NC	DI = North Carolina
DI_VA	AI = Virginia
YRS_SNC_COLLEGE1	Attending college
YRS_SNC_COLLEGE2	Graduated less than 1 year ago
YRS_SNC_COLLEGE3	Graduated 1–3 years ago
YRS_SNC_COLLEGE4	Graduated 4–6 years ago
YRS_SNC_COLLEGE5	Graduated 7–10 years ago
YRS_SNC_COLLEGE6	Graduated 10+ years ago
race1	Asian
race2	African American
race3	Hispanic
race4	White
race5	Other
race6	Unspecified

Variable	Description
Ebest1	Best language is English
Ebest2	Best language is other than English or Spanish
Ebest3	Best language is Spanish
First_Lang_E	First language is English only
First_Lang_EO	First language is English and another language (bilingual)
First_Lang_O	First language is other than English only
P_Lang_E	Proficient in English only
P_Lang_ES	Proficient in English and Spanish
P_Lang_EO	Proficient in English and other language (non-Spanish)
ED_Lvl_1	College freshman or sophomore
ED_Lvl_2	College junior
ED_Lvl_3	College senior
ED_Lvl_4	Bachelor's degree
ED_Lvl_5	Bachelor's degree + courses
ED_Lvl_6	Master's degree
ED_Lvl_7	Master's degree +
UGPA1	Undergraduate GPA: 3.5 – 4
UGPA2	Undergraduate GPA: 3.0 – 3.49
UGPA3	Undergraduate GPA: 2.5 – 2.99
UGPA4	Undergraduate GPA: <2.5
Umajor0	Undergraduate major: middle school education
Umajor1	Undergraduate major: education subject
Umajor2	Undergraduate major: elementary education
Umajor3	Undergraduate major: humanities
Umajor4	Undergraduate major: math and science
Umajor5	Undergraduate major: nonteaching education
Umajor6	Undergraduate major: social sciences
Umajor7	Undergraduate major: special education
Umajor8	Undergraduate major: vocational
Umajor9	Undergraduate major: undecided
Gmajor0	Graduate major: middle school education
Gmajor1	Graduate major: education subject
Gmajor2	Graduate major: elementary education
Gmajor3	Graduate major: humanities
Gmajor4	Graduate major: math and science
Gmajor5	Graduate major: nonteaching education
Gmajor6	Graduate major: social sciences
Gmajor7	Graduate major: special education
Gmajor8	Graduate major: vocational
Gmajor9	Graduate major: undecided
ED_Pro1	Teacher education program: currently attending
ED_Pro2	Teacher education program: formerly attended
ED_Pro3	Teacher education program: never attended
TEACH_PREP1	Teacher preparation program: undergraduate
TEACH_PREP2	Teacher preparation program: 5th year
TEACH_PREP3	Teacher preparation program: master's degree
TEACH_PREP4	Teacher preparation program: alternate route
TEACH_PREP5	Teacher preparation program: other
TEACH_State1	Teach in the same state as testing: yes
TEACH_State2	Teach in the same state as testing: no
Teach_status1	Teaching status: plan to enroll/enrolled in teacher prep program
Teach_status2	Teaching status: recently graduated and will begin teaching soon
Teach_status3	Teaching status: 1 – 3 years of teaching experience
Teach_status4	Teaching status: >3 years of teaching experience
Teach_status5	Teaching status: not planning to teach at this time
Teach_area1	Geographic area to teach: urban
Teach_area2	Geographic area to teach: rural
Teach_area3	Geographic area to teach: suburban
Teach_area4	Geographic area to teach: not planning to teach next year

Note. AI = attending institution. DI = designated institution. GPA = grade point average.

Appendix B

Statistical Results

Table B1 Test 1 Standardized Mean Differences on Covariates in Propensity Score Weighting

	Standardized mean difference	
Variable	Unweighted	Weighted
age	.16	.03
race2	.08	.01
Ebest3	.08	.04
Teach_area4	.06	.01
race3	.06	.01
First_Lang_O	.06	.02
Teach_status4	.05	.01
TEACH_State1	.05	.00
P_Lang_ES	.05	.01
race5	.05	.01
Umajor2	.05	−.01
First_Lang_EO	.04	.02
YRS_SNC_COLLEGE6	.04	.01
TEACH_PREP5	.04	.01
YRS_SNC_COLLEGE5	.04	.01
P_Lang_EO	.04	.03
UGPA3	.04	.01
Teach_status5	.04	.01
ED_Lvl_2	.03	.00
Gmajor1	.03	.02
Umajor1	.03	.00
ED_Lvl_5	.03	.02
Ebest2	.02	.00
Teach_area2	.02	−.01
Gmajor2	.02	−.01
Umajor9	.01	.01
race1	.01	.00
Teach_area1	.01	.00
ED_Pro1	.01	.00
UGPA2	.01	.00
ED_Lvl_1	.01	.00
ED_Pro3	.01	.01
UGPA4	.01	.00
ED_Lvl_4	.01	.01
race6	.00	−.01
TEACH_PREP1	.00	−.02
Umajor0	.00	.00
ED_Lvl_6	.00	.00
Gmajor4	.00	.00
TEACH_PREP3	.00	.01
Umajor8	−.01	.00
Gmajor3	−.01	.01
Umajor7	−.01	.01
YRS_SNC_COLLEGE3	−.01	.00
Teach_status2	−.01	.00
YRS_SNC_COLLEGE1	−.02	−.01
YRS_SNC_COLLEGE4	−.02	.00
Gmajor5	−.02	−.01
Gmajor6	−.02	.00
female	−.02	−.01
Gmajor8	−.02	−.01

Table B1 Continued

Variable	Standardized mean difference	
	Unweighted	Weighted
TEACH_PREP4	−.02	.01
Umajor3	−.02	.00
Umajor5	−.02	−.01
Gmajor9	−.02	.00
YRS_SNC_COLLEGE2	−.02	.00
ED_Lvl_7	−.03	−.01
ED_Pro2	−.03	−.01
TEACH_PREP2	−.03	.00
Teach_status1	−.03	−.01
Teach_status3	−.03	−.01
UGPA1	−.04	.00
Umajor4	−.04	.00
Gmajor7	−.04	−.01
ED_Lvl_3	−.04	−.02
P_Lang_E	−.05	−.01
Ebest1	−.05	−.02
First_Lang_E	−.06	−.01
Umajor6	−.06	.00
TEACH_State2	−.07	.00
Teach_area3	−.10	−.01
race4	−.13	−.02

Table B2 Test 1 Weighted Test Outcome by Test Center and Remote Proctoring (TC − RP)

	Mean scale score difference		Pass rate difference (%)	
	Unweighted	Weighted	Unweighted	Weighted
Total group	−2.01**	−0.56	−3.63**	−1.03
Gender				
Female	−2.01**	−0.55	−3.65**	−1.06
Male	−1.85	−1.00	−2.46	0.05
Race/ethnicity				
White	−0.46	−0.08	−0.82	−0.43
African American	−1.88*	−0.00	−3.37	0.51
Hispanic	−3.39**	−1.57	−2.79	0.99

*Unweighted difference statistically significant at $p < .05$. **Unweighted difference statistically significant at $p < .01$.

Table B3 Test 2 Standardized Mean Differences on Covariates in Propensity Score Weighting

Variable	Standardized mean difference	
	Unweighted	Weighted
Teach_status1	.11	−.01
TEACH_PREP1	.11	−.01
YRS_SNC_COLLEGE1	.09	−.01
Teach_area4	.09	.01
Umajor2	.09	−.01
ED_Lvl_3	.08	.00
ED_Pro1	.07	.00
race1	.06	.02
First_Lang_O	.05	.01
UGPA1	.05	.00
ED_Lvl_2	.04	−.01
Ebest2	.04	.01
P_Lang_EO	.03	.01
TEACH_PREP5	.03	.01
race6	.03	.01
Umajor8	.03	.01
ED_Lvl_1	.02	−.01

Table B3 Continued

Variable	Standardized mean difference	
	Unweighted	Weighted
Gmajor9	.02	−.01
First_Lang_EO	.01	.00
Gmajor4	.01	.01
TEACH_State1	.01	.00
YRS_SNC_COLLEGE5	.01	.01
YRS_SNC_COLLEGE6	.01	.01
Gmajor8	.01	.01
P_Lang_ES	.01	.01
Gmajor1	.01	.01
Teach_status5	.00	.00
Ebest3	.00	.00
Umajor0	.00	.00
race3	.00	.00
Umajor9	.00	.00
Gmajor0	.00	−.01
Teach_area2	.00	.00
Umajor1	−.01	.00
Umajor7	−.01	.00
ED_Pro3	−.01	.00
race2	−.01	.01
TEACH_PREP2	−.01	.00
race4	−.02	−.01
age	−.02	.01
Gmajor3	−.02	.00
TEACH_State2	−.02	.00
female	−.02	−.01
P_Lang_E	−.02	−.01
YRS_SNC_COLLEGE4	−.02	.00
race5	−.02	.00
Gmajor5	−.02	.00
Gmajor6	−.02	.00
UGPA3	−.03	.00
Umajor4	−.03	−.01
Umajor5	−.03	.00
Gmajor2	−.03	−.01
UGPA2	−.03	.00
UGPA4	−.04	.00
Teach_status4	−.04	.00
Teach_area3	−.04	−.01
Ebest1	−.04	−.01
Teach_status2	−.04	.00
Teach_area1	−.04	.00
YRS_SNC_COLLEGE3	−.04	.00
Umajor3	−.05	.00
First_Lang_E	−.05	−.01
ED_Lvl_5	−.05	.00
Gmajor7	−.05	−.01
ED_Lvl_6	−.06	.00
ED_Lvl_4	−.06	.00
ED_Lvl_7	−.07	.00
TEACH_PREP4	−.07	.00
ED_Pro2	−.08	.00
Umajor6	−.09	.00
TEACH_PREP3	−.10	.00
YRS_SNC_COLLEGE2	−.11	−.01
Teach_status3	−.11	.00

Table B4 Test 2 Weighted Test Outcome by Test Center and Remote Proctoring (TC − RP)

	Mean scale score difference		Pass rate difference (%)	
	Unweighted	Weighted	Unweighted	Weighted
Total group	0.40**	0.57++	0.47	0.77
Gender				
Female	0.49**	0.64++	0.71	1.00+
Male	−0.16	0.08	−1.10	−0.86
Race/ethnicity				
White	0.76**	0.93++	1.62**	1.85++
African American	−1.11	−0.84	−4.26*	−0.34
Hispanic	−0.45	0.10	−1.41	−0.52

*Unweighted difference statistically significant at $p < .05$. **Unweighted difference statistically significant at $p < .01$. +Weighted difference statistically significant at $p < .05$. ++Weighted difference statistically significant at $p < .01$.

Table B5 Test 3 Standardized Mean Differences on Covariates in Propensity Score Weighting

	Standardized mean difference	
Variable	Unweighted	Weighted
race1	.07	.04
P_Lang_EO	.07	.03
TEACH_PREP1	.07	−.01
age	.07	.04
Teach_area2	.07	.01
Gmajor0	.05	.03
ED_Pro1	.04	.00
ED_Lvl_2	.04	.00
Umajor2	.04	.02
ED_Lvl_1	.04	.01
Umajor1	.04	−.01
First_Lang_EO	.03	.01
YRS_SNC_COLLEGE3	.03	.01
Umajor0	.03	.01
Teach_status3	.03	.01
Gmajor7	.03	.02
P_Lang_ES	.03	.02
YRS_SNC_COLLEGE1	.02	−.02
Ebest2	.02	.02
First_Lang_O	.02	.00
Teach_area4	.02	−.01
UGPA3	.01	.01
Ebest3	.01	.00
ED_Lvl_7	.01	.02
Teach_status5	.01	.00
race3	.00	−.01
TEACH_State2	.00	−.01
Teach_status1	.00	−.01
YRS_SNC_COLLEGE4	.00	.01
UGPA2	.00	.01
TEACH_PREP5	.00	.01
Umajor8	.00	.00
Gmajor9	.00	.01
race2	.00	.01
Gmajor6	.00	−.02
Umajor6	.00	.00
YRS_SNC_COLLEGE6	.00	.01

Table B5 Continued

Variable	Standardized mean difference	
	Unweighted	Weighted
Ebest1	.00	.00
Gmajor5	.00	.02
race6	−.01	−.01
UGPA1	−.01	−.02
ED_Lvl_4	−.01	.00
ED_Lvl_5	−.01	.00
Umajor4	−.01	.02
ED_Lvl_3	−.01	−.02
ED_Pro2	−.01	.00
Teach_status2	−.01	−.01
TEACH_PREP3	−.01	.01
race5	−.01	−.01
Gmajor4	−.01	.00
race4	−.02	−.01
Teach_area1	−.02	.01
Gmajor2	−.02	.01
Gmajor3	−.02	.00
ED_Lvl_6	−.02	.01
TEACH_State1	−.02	−.01
Gmajor1	−.03	−.02
Umajor7	−.03	.00
TEACH_PREP2	−.03	.00
P_Lang_E	−.03	−.02
First_Lang_E	−.03	.00
Gmajor8	−.03	−.01
Teach_status4	−.03	.01
Umajor3	−.04	−.01
ED_Pro3	−.04	.01
Umajor5	−.04	−.01
YRS_SNC_COLLEGE5	−.04	−.01
UGPA4	−.04	−.01
female	−.04	−.01
YRS_SNC_COLLEGE2	−.04	−.01
Teach_area3	−.06	−.02
TEACH_PREP4	−.08	−.01
Umajor9	−.10	.01

Table B6 Test 3 Weighted Test Outcome by Test Center and Remote Proctoring

	Mean scale score difference		Pass rate difference (%)	
	Unweighted	Weighted	Unweighted	Weighted
Total group	−1.05*	−0.93+	−1.75	−1.68
Gender				
Female	−0.75	−0.66	−0.37	−0.34
Male	−1.19*	−1.10	−2.53	−0.26
Race/ethnicity				
White	−0.52	−0.46	0.02	−0.14
African American	−4.37**	−4.02++	−13.27**	−12.65++
Hispanic	−2.05	−0.82	−2.27	0.62

*Unweighted difference statistically significant at $p < .05$. **Unweighted difference statistically significant at $p < .01$. +Weighted difference statistically significant at $p < .05$. ++Weighted difference statistically significant at $p < .01$.

Table B7 Test 3, African American, Standardized Mean Differences in Propensity Score Weighting

	Standardized mean difference	
Variable	Unweighted	Weighted
DI_VA	.29	.06
age	.23	.17
AI_VA	.22	.04
ED_Lvl_5	.18	.10
ED_Pro1	.18	.01
Teach_area2	.17	.00
YRS_SNC_COLLEGE3	.14	.05
Gmajor7	.10	.00
AI_NC	.10	.02
Gmajor9	.09	−.03
TEACH_State2	.08	.05
Umajor1	.08	−.03
TEACH_PREP1	.08	.02
Teach_status1	.08	.00
First_Lang_E	.07	−.03
P_Lang_EO	.07	.07
DI_NC	.06	−.04
TEACH_PREP5	.06	.05
Gmajor0	.06	.08
ED_Lvl_7	.04	.08
Gmajor5	.04	.04
YRS_SNC_COLLEGE1	.04	.04
First_Lang_EO	.04	.06
DI_AR	.04	−.02
UGPA3	.03	.00
Umajor4	.03	.02
ED_Lvl_3	.03	.02
P_Lang_ES	.03	.01
ED_Lvl_1	.03	.07
Teach_status2	.02	.00
AI_AR	.01	.00
Teach_status4	.01	.04
Umajor6	.01	.07
YRS_SNC_COLLEGE5	−.01	−.05
ED_Lvl_2	−.01	−.03
Umajor8	−.01	−.09
UGPA2	−.01	.03
Teach_area3	−.01	.02
YRS_SNC_COLLEGE4	−.02	−.03
TEACH_PREP2	−.02	.00
Umajor0	−.02	−.05
TEACH_PREP3	−.02	−.04
Umajor2	−.03	.02
Umajor7	−.03	.03
UGPA4	−.03	−.07
ED_Lvl_6	−.04	−.01
UGPA1	−.04	−.02
P_Lang_E	−.04	−.03
Umajor5	−.05	.00
ED_Pro2	−.05	.04
Umajor9	−.05	−.01
TEACH_State1	−.08	−.04
YRS_SNC_COLLEGE6	−.08	.03
TEACH_PREP4	−.10	−.04

Table B7 Continued

Variable	Standardized mean difference	
	Unweighted	Weighted
Teach_area1	−.11	−.01
Gmajor1	−.11	.01
ED_Lvl_4	−.11	−.11
female	−.11	.02
Teach_status3	−.12	−.04
Gmajor3	−.12	−.18
ED_Pro3	−.12	−.04
AI_NA	−.14	−.02
Teach_area4	−.15	−.01
AI_LA	−.16	.02
YRS_SNC_COLLEGE2	−.17	−.06
Umajor3	−.18	−.08
Gmajor2	−.19	.00
Gmajor6	−.19	−.05
DI_LA	−.45	−.04

Table B8 Total Group Regression of Scale Score on Covariates: R^2 (Adjusted R^2)

	No. predictors	Total group	Test center	Remote proctoring
Test 1	72	.286 (.280)	.317 (.304)	.268 (.259)
Test 2	72	.175 (.172)	.190 (.185)	.173 (.168)
Test 3	72	.257 (.242)	.320 (.285)	.241 (.216)

Table B9 Total Group Logistic Regression of PASS/FAIL on Covariates: Percentage Concordant

	No. predictors	Total group	Test center	Remote proctoring
Test 1	72	81.6	82.5	81.7
Test 2	72	73.5	75.5	73.0
Test 3	72	76.7	80.4	76.2

Table B10 Total Group Logistic Regression of Test Center/Remote Proctoring on Covariates: Percentage Concordant

	No. predictors	Total group
Test 1	66	58.6
Test 2	72	57.0
Test 3	55	57.9

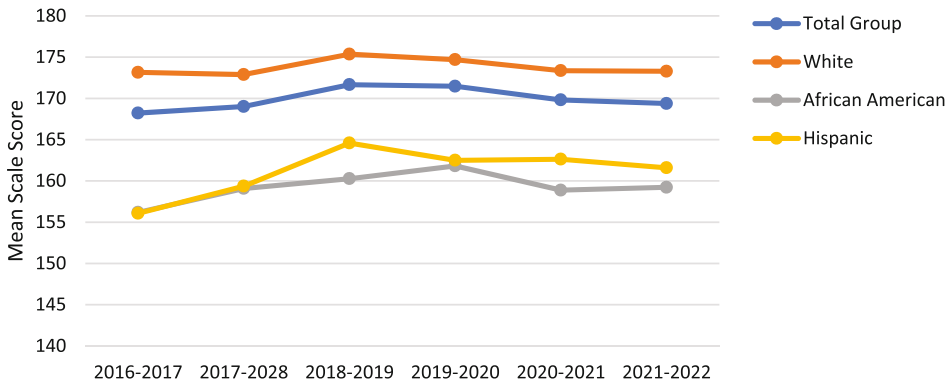

Figure B1 Mean test score by year for Test 1 (first-time test takers).

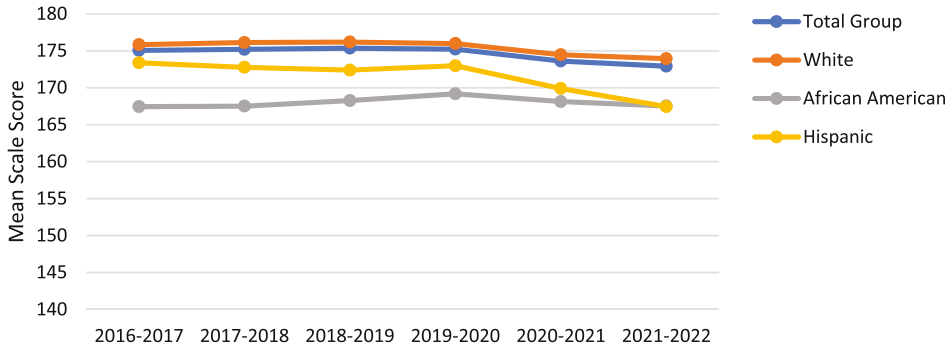

Figure B2 Mean test score by year for Test 2 (first-time test takers).

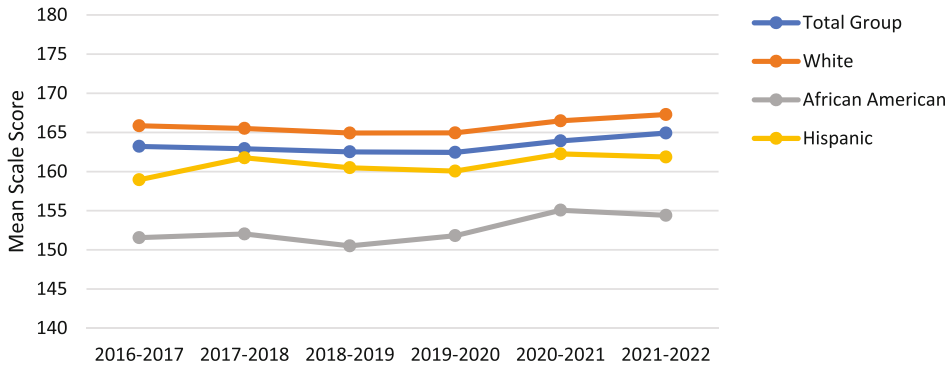

Figure B3 Mean test score by year for Test 3 (first-time test takers).

Suggested citation:

Miao, J., Cao, Y., & Walker, M. E. (2024). *Detecting the impact of remote proctored at-home testing using propensity score weighting* (Research Report No. RR-24-11). ETS. https://doi.org/10.1002/ets2.12386

Action Editor: Daniel F. McCaffrey

Reviewers: Hongwen Guo and Sandip Sinharay