Check for updates

Research Report

Does Retest Effect Impact Test Performance of Repeaters in Different Subgroups?

ETS RR–20-16

Jiawen Zhou
Yi Cao

December 2020

Discover this journal online at
Wiley Online Library
wileyonlinelibrary.com

ETS Research Report Series

EIGNOR EXECUTIVE EDITOR

John Mazzeo
Distinguished Presidential Appointee

ASSOCIATE EDITORS

Beata Beigman Klebanov
Senior Research Scientist

Heather Buzick
Senior Research Scientist

Tim Davey
Research Director

John Davis
Research Scientist

Marna Golub-Smith
Principal Psychometrician

Priya Kannan
Managing Research Scientist

Sooyeon Kim
Principal Psychometrician

Anastassia Loukina
Senior Research Scientist

Gautam Puhan
Psychometric Director

Jonathan Schmidgall
Research Scientist

Jesse Sparks
Research Scientist

Michael Walker
Distinguished Presidential Appointee

Rebecca Zwick
Distinguished Presidential Appointee

PRODUCTION EDITORS

Kim Fryer
Manager, Editing Services

Ayleen Gontz
Senior Editor

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

Does Retest Effect Impact Test Performance of Repeaters in Different Subgroups?

Jiawen Zhou & Yi Cao

Educational Testing Service, Princeton, NJ

In this study, we explored retest effects on test scores and response time for repeaters, examinees who retake an examination. We looked at two groups of repeaters: those who took the same form twice and those who took different forms on their two attempts for a certification and licensure test. Scores improved over the two test attempts, and repeaters taking the same form twice tended to have larger score gains than those taking different forms. This trend was more salient for female examinees than male and for White and Hispanic/Latino examinees than examinees from other ethnicity groups. In addition, over a third of the examinee responses were incorrect on the same form for both attempts, indicating that the knowledge required for these items was not addressed after these repeaters failed their initial attempt to answer those items correctly. Factors that may be related to these findings are discussed.

Keywords Retest effect; gender and ethnicity subgroups; certification and licensure assessment; repeaters

doi:10.1002/ets2.12300

Certification and licensure tests are widely employed in a variety of professional fields, such as accounting, medicine, psychology, and teaching, and are taken by hundreds of thousands of candidates every year. The purpose of certification and licensure tests is to assure the public that individuals who practice an occupation or profession have met certain standards (American Educational Research Association [AERA] et al., 2014). Because the results of these tests can have a long-lasting impact on an individual's career, most certification and licensure testing programs provide examinees who fail the test with an opportunity to retake the test after a certain blackout period. Such a retest policy is an important practice to ensure fairness because examinees might not fully represent their true level of proficiency on the construct of interest on the initial test attempt. Retest policy is also consistent with professional standards and government guidelines (AERA et al., 2014; Society for Industrial and Organizational Psychology, 2003).

Rates for repeaters, examinees who retake a test, vary among different certification and licensure tests, with repeater rates reaching as high as 50%. A large body of literature on retest effect indicates that repeaters gain higher scores on their second attempt than on their initial attempt (Boulet et al., 2003; Feinberg et al., 2015; Geving et al., 2005; Raymond et al., 2009). These score gains can be attributed at least partly to several factors, including true improvement on the construct being measured, reduction in test anxiety and increase in test-taking skills, and memorization of previous test content (Lievens et al., 2005). To prevent repeat candidates from obtaining unwarranted score gains due to memorization of previous test content, most high-stakes certification and licensure testing programs develop multiple forms of a test that are based on detailed test specifications and thus are expected to be parallel in content and difficulty level. However, the cost and technical difficulty associated with developing and administering multiple forms pose challenges that certification and licensure testing programs cannot overlook. For most high-stake assessments, producing new test items requires a series of steps including writing, reviewing, pilot testing, analysis, and revision. Resolution for each step requires professional engagement and a substantial budget.

To resolve this dilemma, a growing body of research on retest effects has been conducted within the context of certification and licensure tests to investigate whether repeaters truly benefit from seeing the same items or a same form twice. Boulet et al. (2003) compared the performance of candidates who graduated from medical schools on a clinical skills assessment for licensing physicians and found that repeaters generally scored higher on retest than their initial attempt, but the results showed no benefit of seeing the same patient/clinical problem twice. In their study, Geving et al. (2005) included more than 9,000 participants who took a real estate licensing examination. The examination was computer-based

Corresponding author: J. Zhou, E-mail: jzhou@ets.org

and each administration contained 80 multiple-choice items from an item bank with 643 items. On average, the different forms shared 12% of the items in common. The results indicated that the average change in score between the first and second administrations was an increase of 0.62 standard deviation (*SD*) units. Gevin et al. also analyzed score gains on the common items that appeared in both administrations. When repeaters received an identical item in two administrations, the answer changed from correct to incorrect as often as from incorrect to correct. There was no advantage in seeing the same items on the second administration.

Beside the studies investigating the benefit to repeaters of seeing a small number of common items in different forms, there were also studies on the performance of examinees who took the same test form twice. Raymond et al. (2007) compared repeaters' performance on two certification tests in computed tomography and radiography. They discovered that 39 examinees who received the same form in their second administration of the computed tomography test had an increase of 0.80 *SD* units in score, and 41 examinees who received a different form in their retake increased scores by 0.78 *SD* units. This trend was also true for the radiography test. A total of 102 repeaters taking the same form on their second administration gained 0.47 *SD* units and 663 examinees who repeated the test by taking a different form increased scores by 0.48 *SD* units. On both tests, the average score gain was almost same for examinees taking the same form twice as for examinees taking different test forms.

Raymond et al. (2009) conducted a follow-up experimental study of 541 examinees who failed a national certification test in radiography. On the second attempt, these repeaters were randomly assigned to receive either the same form or a different form of the test. Even though the same-form group had a shorter response time on the retest, the difference on score gains for both groups was negligible. However, a widely cited meta-analysis study regarding retest effects by Kulik et al. (1984) indicated that repeaters' average score gain on the same form was 0.42 *SD* units, but only 0.23 *SD* units on different forms. Of the 40 studies from which Kulik et al. collected data, the majority involved aptitude tests. This finding was confirmed by another meta-analysis study focusing on cognitive ability tests conducted by Hausknecht et al. (2007).

Evidence from the previous studies reached only one clear conclusion: There are consistent score gains due to retesting. Other findings regarding retest effects are far from conclusive. In addition, these existing studies were mostly limited to the medical field, with some research involving experimental designs, which makes it hard to generalize the findings to other professional fields. In our study, we gathered empirical evidence about retest effects in an authentic licensure testing environment outside of the medical field. Our goal was to examine retest effects on test scores and response time for repeaters taking the same form twice and for those taking different forms on their first and second attempts. We also studied retest effects in different gender and ethnicity subgroups. In addition, we explored how repeaters taking the same form twice changed their item responses over the two attempts. The following specific research questions are addressed in the present study:

- Research Question 1: Is the time interval between repeaters' initial and second test attempts associated with the repeaters' test score change? Is it associated with the repeaters' test response time change?
- Research Question 2: How much do repeaters' test scores and response times change between their initial and second attempts? How much do the changes differ between repeaters taking the same form twice and repeaters taking different forms?
- Research Question 3: How much do the answers to Research Question 2 differ for repeaters as a function of their gender and ethnicity subgroup membership?
- Research Question 4: When repeaters took the same form of the test a second time, how many items did they answer correctly the second time but not the first time, and vice versa?

Method

Data

Data from a large-scale computer-based licensure test for entry-level mathematics educators were used. The multiple-choice test lasts 150 minutes and consists of 50 scored items and 10 unscored pretest items. The reported scale score has possible range of 100–200. This test has an average reliability of 0.88 with a raw score standard error of measurement of 3.0. Multiple test forms were assembled based on a detailed test blueprint and statistical properties. Each form shared a certain number of common items with a particular (i.e., reference) form for equating purpose but shares almost no items with any other forms. Twelve administrations are scheduled for this test annually, with one form per administration and multiple forms in rotation. Examinees who fail the first attempt can retest in 21 days, but no limitations are in place concerning the number of retest attempts. Given that one form is administered in each testing window, repeaters might receive the same form, a form containing shared common items, or a completely different form at their retests.

A total of 26,319 individual examinees took the test from October 2013 to December 2017; of these examinees, 9,124 (approximately 35%) retook the test. Among these repeaters, 1,081 repeaters received the same form on the second attempt, and 4,827 repeaters received a different form. The remaining 3,216 repeaters received a form that shared common items with the form they had taken at the initial attempt. Based upon the research purpose, the final sample for analyses included repeaters who took either the same form ($n = 1{,}081$) or a completely different form ($n = 4{,}827$) on their second attempt. The time interval between these repeaters' two test attempts varied from 22 to 1,443 days, with a mean of 129 days and an *SD* of 156 days.

Analyses

Research Question 1

To answer Research Question 1, the sample was divided into 16 groups based on the time between the initial and second attempts (e.g., fewer than 30 days, 31 – 60 days, 61 – 90 days). Repeaters who took the second test more than 450 days after the initial test, which are more than 2 *SD*s, were grouped together. A histogram is used to show the frequency distribution of repeaters by the retest time interval. Repeaters' test score change (on a scale score) and test response time change (in minutes) across different retest time intervals were then compared using separate box and whisker plots. Each box and whisker plot depicts a five-number summary of a dataset: A box encloses the first quartile (Q1) through the third quartile (Q3), that is, the interquartile range (IQR). A horizontal line divides the box at the median. The lower whisker extends from Q1 to the smallest data value greater than or equal to Q1 – 1.5*IQR, while the upper whisker extends from Q3 to the largest data value less than or equal to Q3 + 1.5*IQR.[1] Outliers, defined as values that are outside of the upper and lower quartiles by at least 1.5 times the IQR, are excluded from the plot.

Research Question 2

To answer Research Question 2, descriptive statistics as well as effect size using Cohen's *d* for repeated measures (Cohen, 1988) was first calculated to show how large the difference was in changes in scores and response times over two test attempts. These analyses were conducted for both the total group and for two different groups of repeaters: the repeaters who received a same form on their first and second attempts (same-form repeaters) and the repeaters who received a different form on their second attempt (different-form repeaters). Box and whisker plots were then drawn to present these differences in scores and response times between the two attempts.

Research Question 3

Two demographic group memberships were considered in this study: gender with two subgroups (male and female) and ethnicity with four subgroups (White, African American, Asian American, and Hispanic or Latino). First, descriptive statistics and effect size were calculated to show the magnitude of difference between the test scores or between the response times for the first and second test attempts. Second, whether each group took the same or different forms over two attempts was considered for each of the two demographic groups (e.g., female repeaters taking same forms over two attempts). Effect size was calculated for each combination. Box and whisker plots were then graphed.

Research Question 4

To answer this question, we first compared each individual repeater's 0/1 scores between the initial and second attempts for each item and categorized item-level score changes into four scenarios: correct to correct, correct to incorrect, incorrect to correct, and incorrect to incorrect. Then, across all items and all repeaters, we summed the number of these score changes in each of the four scenarios and computed the percentages for each scenario. These analyses were conducted on only two forms, both with more than 100 same-form repeaters.

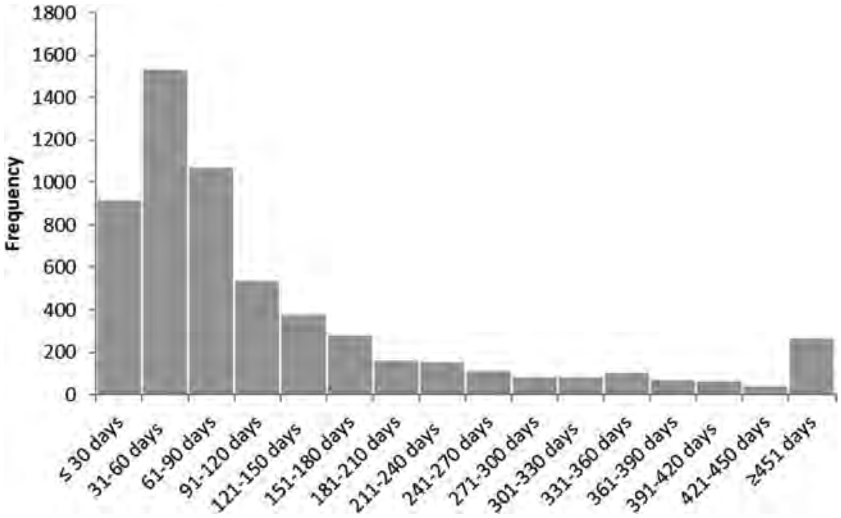

Figure 1 Repeaters' frequency distribution by retest time interval.

Results

Results for Research Questions 1–4 are presented in order. Regarding Research Questions 2 and 3, results focusing on changes in test score are presented first, followed by results for changes in test response time.

Research Question 1

Figure 1 presents the frequency distribution for 5,908 repeaters by retest time interval. The distribution is positively skewed, with approximately 60% of the repeaters retaking the test within 3 months (i.e., 90 days) and more than 80% within half a year (i.e., 180 days) after their initial attempt.

Repeaters' changes in test score (on the scale score) and in response time (in minutes) between the two attempts are illustrated in Figures 2 and 3. Figure 2 shows that across 16 different retest time intervals, changes in individual repeaters' test scores were in both directions, either increasing (above the zero line) or decreasing (below the zero line). However, the means of repeaters' score changes, marked as "x" in the plot, did not differ much, and the IQR of repeaters' score changes differed very little with the length of time interval. For repeaters retaking the test within 180 days, which represented 80% of all repeaters, their mean score changes varied from 5.4 to 6.8 score points. Changes in individual repeaters' response time were in both directions. There was not much variation in the medians of repeaters' response time changes across different retest time intervals (all around 0 minutes), but there were some fluctuations in the means as shown in Figure 3. However, the variation in mean test response time changes of repeaters taking the test a second time across 16 time intervals was less than 6 minutes, with those taking their retest within half a year ranging from 0.2 to 3.2 minutes. This appeared to be simply an effect of sample size. Where the groups were large (within 180 days), the group means did not differ much. After 180 days, the groups got smaller, and their means differed more. It was concluded that retest time interval did not play an active role in changes in repeaters' test scores and test response times in this study. This factor is therefore not considered in the following analyses of the present study.

Test Score Change

Research Question 2

Table 1 presents the summary statistics for test scores of all 5,908 repeaters on their initial and second test attempts. Compared to their initial attempt, examinees improved an average of 6.2 test score points with a larger *SD* on their second test attempt. The correlation between the test scores for both attempts is 0.72. The effect size of the mean test score change between the first and second test attempts is 0.54, indicating a medium effect. Overall repeaters had higher scores in the second attempt.

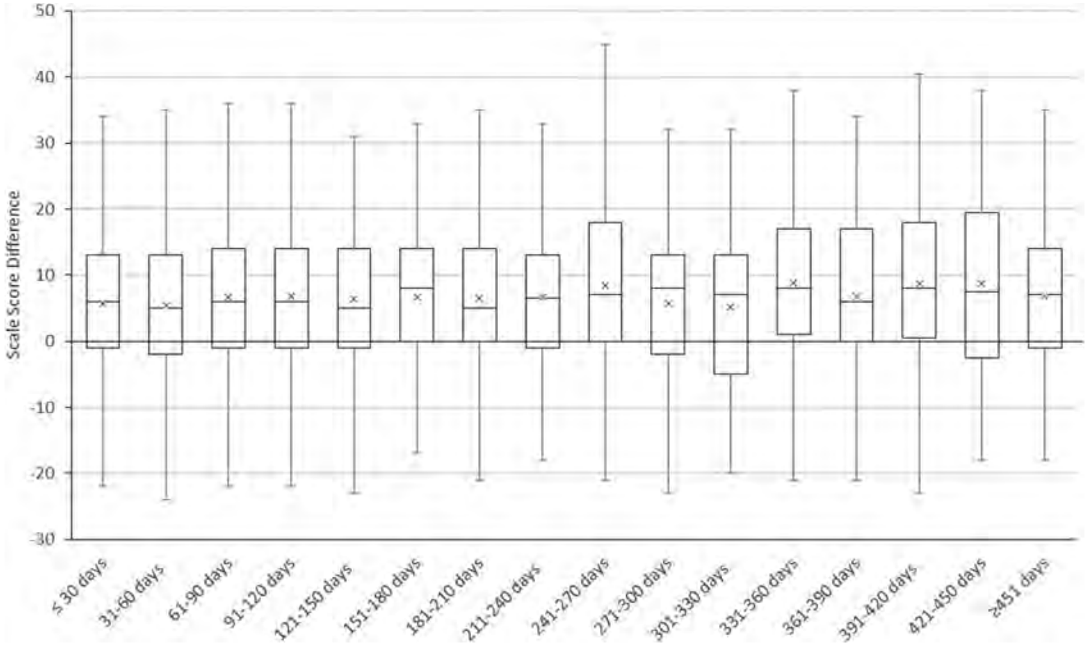

Figure 2 Repeaters' score change between the initial and second attempts.

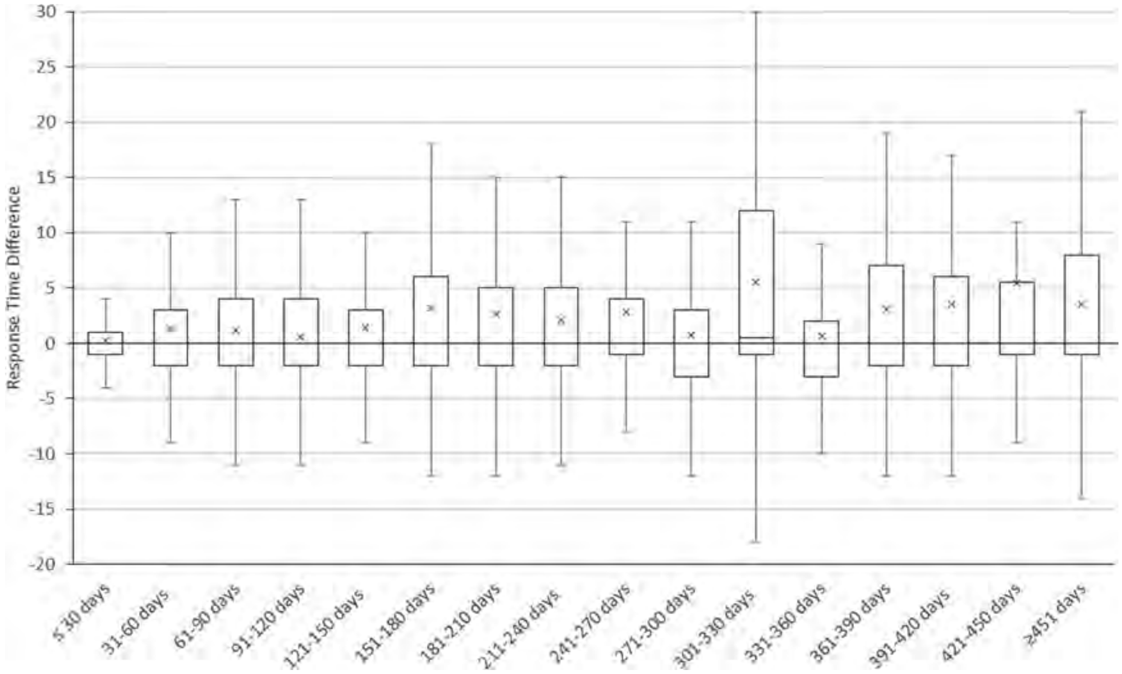

Figure 3 Repeaters' response time change between the initial and second attempts.

Table 2 presents the same type of information shown in Table 1 for the same-form and different-form repeater groups, and Figures 4 and 5 display the results using the box and whisker plots. The mean test scores for the initial and second attempts indicated that, regardless of receiving the same form or different forms, repeaters tended to score higher with a larger variation on their retest. However, same-form repeaters improved more on the second attempt than did different-form repeaters. Same-form repeaters had an average increase of 8.0 points on test score (effect size of 0.74), while different-form repeaters improved 5.9 points (effect size of 0.50). The correlation between the test scores for two attempts is 0.77 for same-form repeaters and 0.71 for different-form repeaters.

Table 1 Summary Statistics for Repeater Test Score on Initial and Second Attempts

	Initial attempt		Second attempt		Score change			
N	Mean	*SD*	Mean	*SD*	Mean	*SD*	Correlation	Effect size
5,908	139.3	13.3	145.6	16.4	6.2	11.6	0.72	0.54

Table 2 Summary Statistics for Same-Form and Different-Form Repeaters

		Initial attempt	Second attempt			Score change			
Form	*N*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Correlation	Effect size
Same	1,081	138.0	13.5	146.0	16.9	8.0	10.8	0.77	0.74
Different	4,827	139.6	13.3	145.5	16.3	5.9	11.7	0.71	0.50

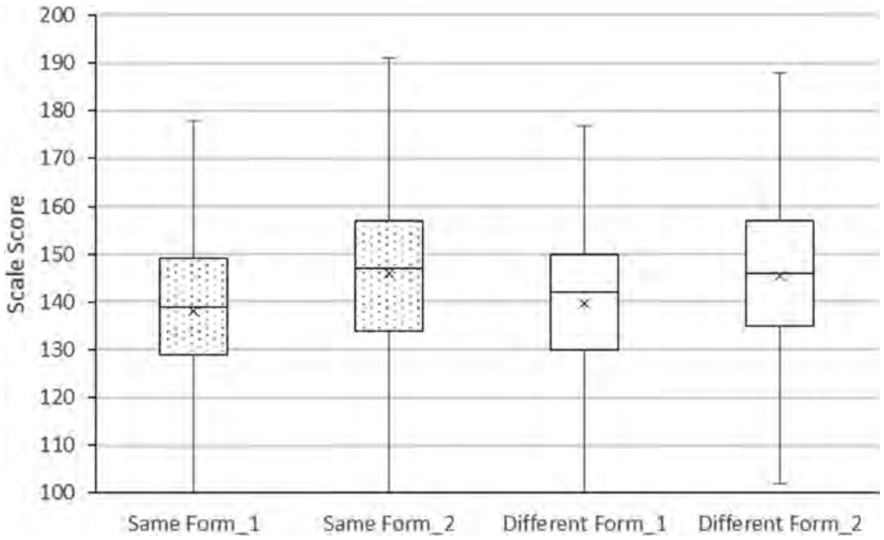

Figure 4 Same-form and different-form repeaters' scores on their initial and second attempts.

Research Question 3

Gender

Table 3 and Figure 6 present the summary statistics and distribution of male and female repeaters' test scores for their two test attempts. In general, repeaters had higher scores on their second attempt. This trend was true for both male and female repeaters. The magnitude of score change was fairly comparable for both males (6.0 as the average score change) and females (6.4 as the average score change) when the particular form (either the same or different form) that they received on the retest was not taken into account for the score change calculation. The correlation between the test scores for both test attempts for both male and female repeaters is 0.72. The effect sizes of score change for male and female subgroups are both above 0.5, as presented in Table 3, indicating a substantial difference.

In addition to the above investigation on the gender main effect, we further examined the potential interaction effect between gender and whether repeaters received the same or a different form during retesting. Table 4 and Figures 7 and 8 present the test score change of male and female repeaters taking the same or a different form on the second attempt. Both male and female same-form repeaters performed better than different-form repeaters. Among repeaters taking the same form twice, female repeaters showed larger score gain (i.e., an average of 8.5 score points) as compared to male repeaters (i.e., 6.9 score points on average), but among repeaters taking different forms for the two attempts, the results were similar for males and female repeaters (i.e., 5.8 score points for males versus 5.9 score points for females). In summary, females benefited more than males from taking the same form twice.

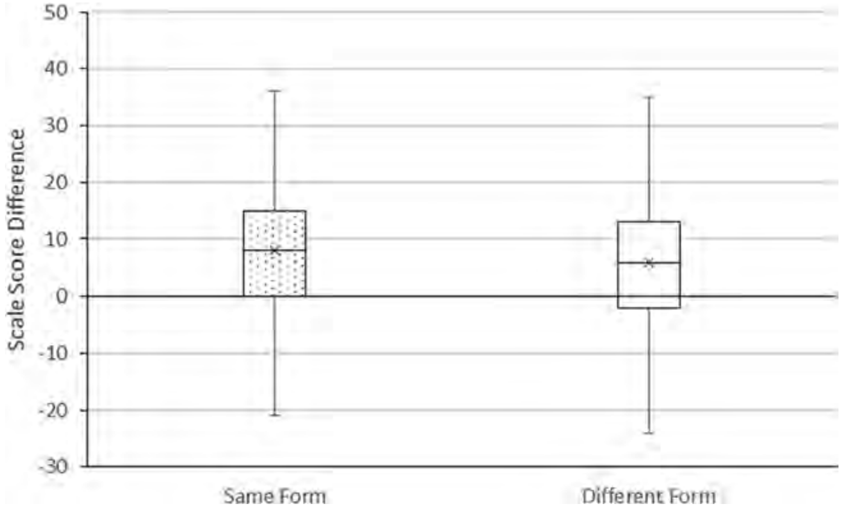

Figure 5 Same-form and different-form repeaters' score changes.

Table 3 Summary Statistics for Repeater Test Scores for Initial and Second Attempts by Gender

		Initial attempt		Second attempt		Score change			
Gender	*N*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Correlation	Effect size
Male	2,092	140.0	13.4	146.0	16.6	6.0	11.6	0.72	0.52
Female	3,816	139.0	13.3	145.3	16.4	6.4	11.5	0.72	0.55

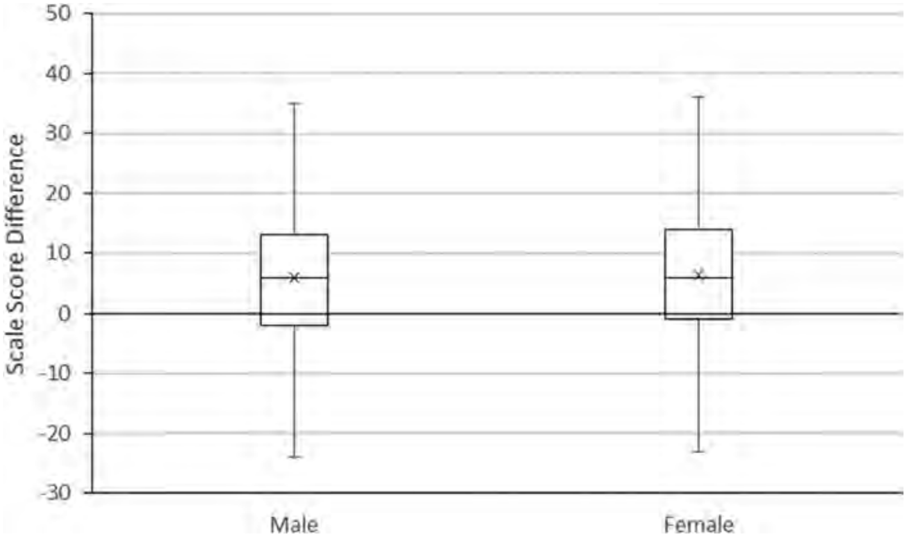

Figure 6 Repeaters' score change between two attempts by gender.

Ethnicity

Table 5 and Figure 9 present the summary statistics and distribution of repeaters' test scores for two test attempts by four different ethnic subgroups. Generally, repeaters from all ethnic subgroups obtained higher scores on the second attempt with larger *SD*s. The average score gain was the highest for Hispanic or Latino repeaters, followed by White and Asian American repeaters, and was the lowest for African American repeaters. The average score gain for African American repeaters, 3.7, was only about half as large as for Hispanic or Latino repeaters, 7.5. The correlation between the test scores on both attempts for each of the four ethnic repeater subgroups ranges from 0.69 to 0.74.

Table 4 Summary Statistics for Repeater Test Score Change by Gender and Form

	Test score change					
Gender	Form	*N*	Mean	*SD*	Correlation	Effect size
Male	Same	371	6.9	11.1	0.75	0.62
	Different	1,721	5.8	11.7	0.71	0.50
Female	Same	710	8.5	10.6	0.78	0.80
	Different	3,106	5.9	11.7	0.70	0.50

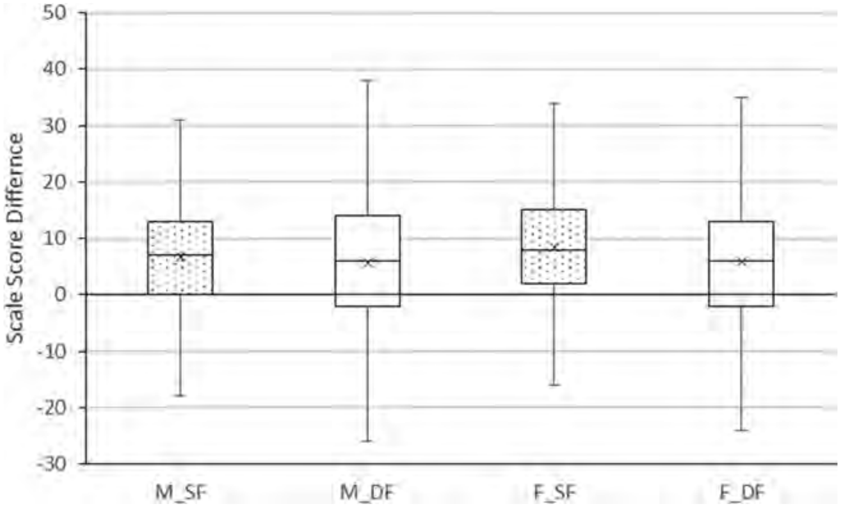

Figure 7 Repeaters' score change between two attempts by gender and form. M_SF = male same-form repeaters; M_DF = male different-form repeaters; F_SF = female same-form repeaters; F_DF = female different-form repeaters.

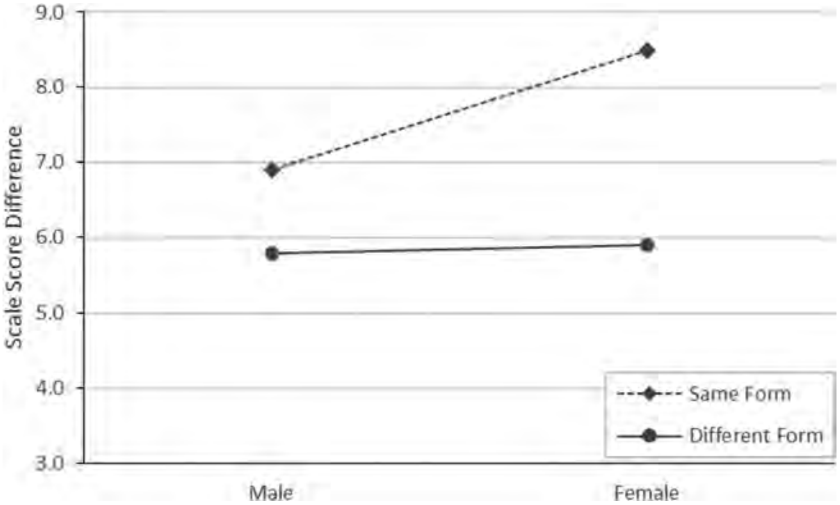

Figure 8 Retest effect by gender and form.

Table 6 and Figure 10 present the test score changes for repeaters across four ethnic subgroups combined with the consideration of the particular form taken on the second attempt (either the same form or different). Generally, for all four ethnic subgroups, the same-form repeaters performed better than the different-form repeaters. Among the same-form repeaters, the White and Hispanic/Latino repeaters improved most, followed by Asian American repeaters, with the African American repeaters showing the smallest score increase. For the different-form repeaters, the Hispanic/Latino repeaters improved the most while the African American improved the least. Figure 11 presents ordinal interaction

Table 5 Summary Statistics for Repeater Test Scores on Initial and Second Attempts by Ethnicity

		Initial attempt		Second attempt		Score change			
Ethnicity	*N*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Correlation	Effect size
White	4,179	140.9	12.5	147.5	15.7	6.6	11.5	0.69	0.57
African American	661	130.5	14.5	134.2	16.2	3.7	11.5	0.72	0.33
Asian American	206	139.4	13.3	145.5	16.6	6.1	11.2	0.74	0.55
Hispanic or Latino	223	135.9	12.9	143.4	16.2	7.5	11.4	0.72	0.66

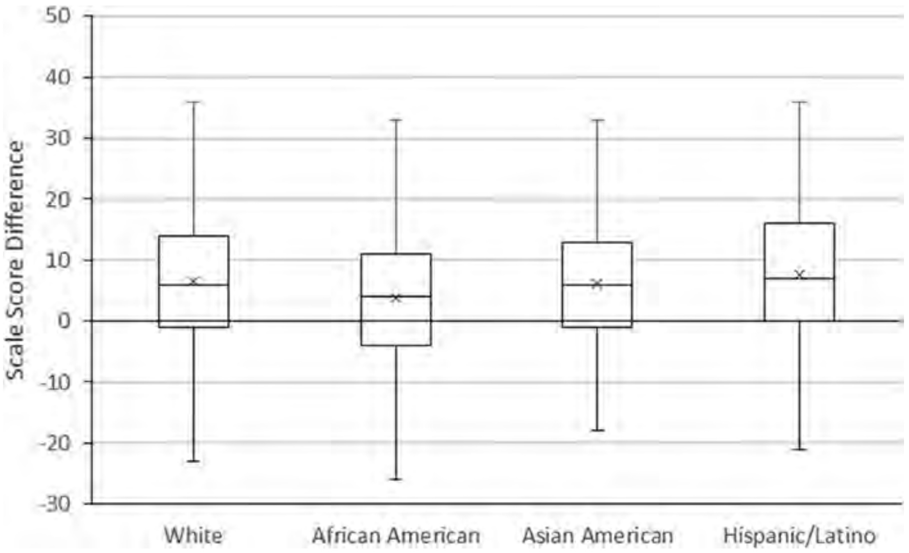

Figure 9 Repeaters' score change between two attempts by ethnicity.

Table 6 Summary Statistics for Repeater Test Score Change by Ethnicity and Form

	Test score change					
Ethnicity	Form	*N*	Mean	*SD*	Correlation	Effect size
White	Same	742	8.5	10.9	0.74	0.78
	Different	3,437	6.2	11.6	0.67	0.53
African American	Same	137	4.5	10.3	0.77	0.43
	Different	524	3.6	11.8	0.71	0.30
Asian American	Same	32	6.8	12.4	0.61	0.55
	Different	174	6.0	11.0	0.76	0.55
Hispanic or Latino	Same	54	8.6	11.1	0.72	0.78
	Different	169	7.2	11.4	0.72	0.63

plots between ethnicity subgroups and different second attempt-forms. In terms of score gains, the difference between same-form and different-form repeaters was greater for the White and Hispanic/Latino subgroups, compared to African American and Asian American repeaters.

Test Response Time Change

Research Question 2

Table 7 provides the summary statistics of 5,908 repeaters' test response time (in minutes) on their initial and second test attempts. It shows repeaters spent an average of 1.5 more minutes (out of 150 minutes in total) on their second test attempt. The mean difference of test response time between the two test attempts is negligible while the large *SD* of individual

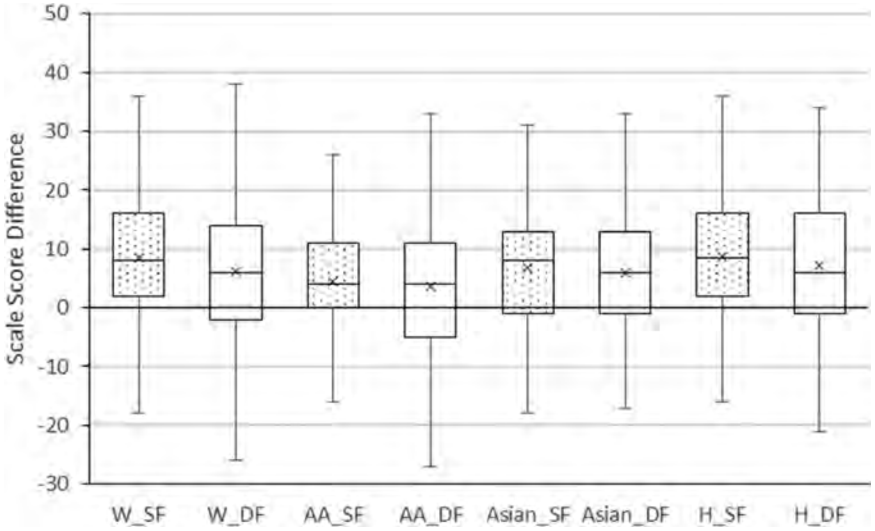

Figure 10 Repeaters' score change between two attempts by ethnicity and form. W_SF = White same-form repeaters; W_DF = White different-form repeaters; AA_SF = African American same-form repeaters; AA_DF = African American different-form repeaters; Asian_SF = Asian American same-form repeaters; Asian_DF = Asian American different-form repeaters; H_SF = Hispanic or Latino same-form repeaters; H_DF = Hispanic or Latino different-form repeaters.

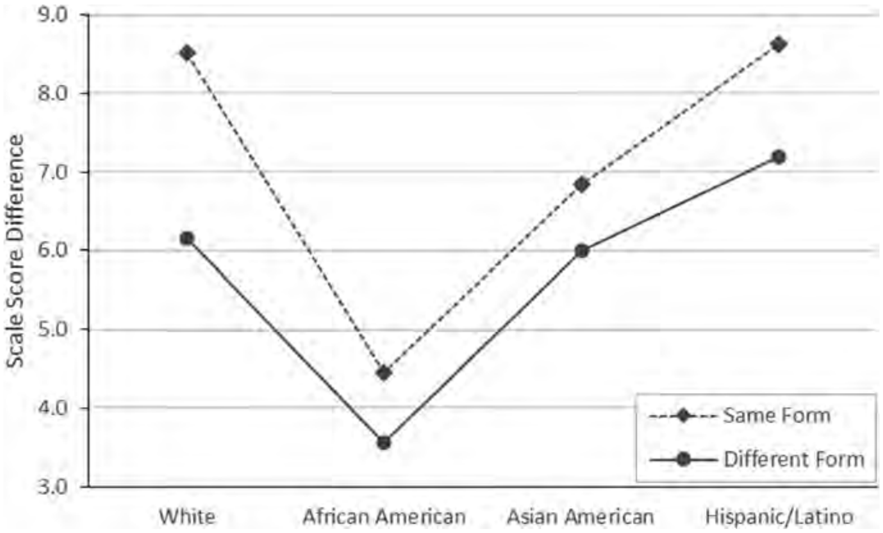

Figure 11 Retest effect by ethnicity and form.

repeater's response time change implies that testing time changed much more for some repeaters than for others. However, there was no systematic pattern in testing time. The correlation between the test response time for the two attempts is 0.59.

Table 8 presents the summary statistics of the time in minutes that same-form versus different-form repeaters took to finish the test on two test attempts. On average, repeaters took slightly more time to complete the test on their second attempt, regardless of whether they received the same or a different form. The same-form repeaters had an increase of 0.8 minutes in test response time while the different-form repeaters had an increase of 1.6 minutes. Therefore, the planned demographic subgroup analyses on test response time change were not conducted.

Research Question 4

To respond to Research Question 4, we focused only on repeaters who received the same form during both initial and second attempts. These same-form repeaters answered each item in the form twice, which provided the opportunity for

Table 7 Summary Statistics for Repeater Response Time (in Minutes) on Initial and Second Attempts

	Initial attempt		Second attempt		Time change			
N	Mean	*SD*	Mean	*SD*	Mean	*SD*	Correlation	Effect size
5,908	139.3	17.1	140.8	15.8	1.5	14.9	0.59	0.10

Table 8 Summary Statistics for Repeater Response Time (in Minutes) on Initial and Second Attempts by Form

		Initial attempt		Second attempt		Time change			
Form	*N*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Correlation	Effect size
Same	1,081	138.9	17.7	139.7	17.4	0.8	15.1	0.63	0.05
Different	4,827	139.4	17.0	141.0	15.4	1.6	14.9	0.58	0.11

Table 9 Score Change Patterns Across All Items in Form 1

		Attempt 2		
		Correct (%)	Incorrect (%)	Total (%)
Attempt 1	Correct	35	10	45
	Incorrect	17	38	55
Total		52	48	100

Note. $N = 761$.

us to investigate changes in 0/1 responses to each item over two attempts. There are four different scenarios: a correct response to an item on both attempts, correct on Attempt 1 but incorrect on Attempt 2, incorrect on Attempt 1 but correct on Attempt 2, and incorrect on two attempts. Tables 9 and 10 present the corresponding percentages of score changes falling into each scenario over all items and across all repeaters. It should be noted that multiple forms were included in the analyses for Research Questions 1 to 3. For Research Question 4, however, the analyses were conducted on only two forms that had more than 100 same-form repeaters. The findings in terms of score change patterns on the two forms are similar to each other. First, repeaters tended to answer more items correctly on their second attempt than on first attempt. On Form 1, repeaters correctly answered 45% of the items on their first attempt but 52% on their second attempt. On Form 2, the percentage of correctly answered items was 49% on the first attempt and 55% on the second. This outcome was not surprising as repeaters' average score increased on the second attempt. Second, 10% of item responses on Form 1 (11% on Form 2) were changed from correct to incorrect across the two test attempts. A possible explanation is that repeaters slipped on the second attempt due to carelessness or other reasons, or they guessed on both attempts but were lucky on only the first attempt. Third, 17% of item responses on both Form 1 and Form 2 changed from incorrect to correct. This can be due to either a lucky guess on the second attempt or actual learning occurred between the two test attempts. Lastly, 38% of the responses on Form 1 and 34% on Form 2 are incorrect on both attempts. When an examinee incorrectly answers the same item twice, it is likely that the examinee has not mastered the required knowledge and skills to solve the item.

Discussion

Existing research on retest effect has been mostly limited to testing in the medical field. The present study added more empirical evidence into existing research literature regarding the retest effect for testing outside of the medical area. This study examined retest effect in terms of test score change and response time change over initial and second attempts between repeater groups receiving the same versus different forms, as well as with the consideration of the effect on gender and ethnicity subgroups. Also, we explored how repeaters change their item-level responses over two attempts. The present study revealed a few findings that are worth further discussion.

Table 10 Score Change Patterns Across All Items in Form 2

		Attempt 2		
		Correct (%)	Incorrect (%)	Total (%)
Attempt 1	Correct	38	11	49
	Incorrect	17	34	51
Total		55	45	100

Note. N = 170.

The primary finding is that repeaters scored higher on their retest compared to their initial attempt, on average. The result of repeaters' score gain over attempts echoes the findings reported by previous research on retest effect in a certification and licensure testing context (Geving et al., 2005; Raymond et al., 2009). Repeaters' test score improvement over attempts might be due to different reasons. One possible explanation is that repeat candidates truly improve on the construct being measured after their first attempt. Testing itself promotes learning, even when different items from the same construct of interest were examined on multiple test attempts (Chan et al., 2006). Examinees might recognize their deficient content topics due to difficulty responding to certain items, which promotes further study with a more clearly focused target. Repeaters might experience less anxiety on the second attempt, which could help them gain better score. In addition, an improvement in general test-taking skills or an increase in familiarity of computer-based testing environment could both contribute to obtain a higher score.

The second noticeable finding is that repeaters receiving a same form on their second attempt tend to have more score gain than those taking different forms. The presence of this phenomenon can probably be attributed to the effect of on-target study as repeaters recognized their deficiencies during the first test attempt. In that case, when they encounter the same items on the second attempt, there is a higher chance that they can correctly answer these items. Apparently, this result is opposite to the findings revealed by Feinberg et al. (2015) and Raymond et al. (2009). In their studies, the mean score gain for same-form and different-form groups was indistinguishable. Therefore, the question of whether administering same form to repeaters would violate score validity is still far from being answered. In addition, the retest effect was reflected in demographic groups considered in this study. Taking the same form seems to benefit repeaters of some subgroups more, in particular, female more than male examinees and White and Hispanic/Latino examines more than other ethnic subgroups.

Another noteworthy finding relates to the score-change pattern over test attempts. Repeaters changed 17% of their answers from incorrect to correct on both Form 1 and Form 2, and at the same time, they changed 10% of their answers from correct to incorrect (11% on Form 2). The difference in these item-level score changes explicates how repeaters' test scores increased on the second attempt. The other notable finding revealed in item-level score-change analyses is that 38% of the responses on Form 1 were incorrect on both occasions (34% on Form 2). An examinee's repeated error is usually attributed to two reasons: The examinee either lacks the knowledge to solve the item or misunderstands the concept being measured by the item (Feinberg et al., 2015). Surprisingly, over one third of examinee responses on the same test form were incorrect on both attempts, suggesting that the knowledge required to correctly answer these items was not achieved after these examinees failed the test on the first attempt. Before retaking the test, if repeaters could identify specific areas of knowledge insufficiencies from their first attempt and look for on-target tutoring or test preparation material, they might effectively remediate those particular knowledge deficits and better prepare themselves for the subsequent retest.

The present study has practical implications and it also leads to some suggestions for future research. First, the pass/fail decision is the most important outcome and use of a certification test. This outcome was not evaluated in this study because not all stakeholders adopt the same cut score. In future studies, decision accuracy and decision consistency indices can be estimated along with pass/fail rates to provide more information on the impact of repeaters on these test quality measures. Second, we could extend the Research Question 4 by further breaking down items as a function of difficulty levels (e.g., easy, medium, hard) to evaluate more deeply the score-change patterns. Third, this study focused on a single test. Future studies can be conducted on tests of different subjects with different item types (e.g., multiple-choice items only versus a combination of multiple-choice items and constructed-response items). We are also interested in the repeater effect on score equating, particularly for small-sample equating (e.g., whether the inclusion of repeaters would impact the results of small-sample score equating or how different percentages of repeaters in the total sample would affect results).

Note

[1] This is the default setting of a box and whisker plot in Microsoft Office 365 ProPlus Excel.

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.

Boulet, J. R., McKinley, D. W., Whelan, G. P., & Hambleton, R. K. (2003). The effect of task exposure on repeat candidate scores in a high-stakes standardized patient examination. *Teaching and Learning in Medicine*, *15*(4), 227–232. https://doi.org/10.1207/S15328015TLM1504_02

Chan, J. C. K., McDermott, K. B., & Roediger, H. L. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General*, *135*(4), 553–571. https://doi.org/10.1037/0096-3445.135.4.553

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge Academic.

Feinberg, R. A., Raymond, M. R., & Haist, S. A. (2015). Repeat testing effects on credentialing exams: Are repeaters misinformed or uninformed? *Educational Measurement: Issues and Practice*, *34*(1), 34–39. https://doi.org/10.1111/emip.12059

Geving, A. M., Webb, S., & Davis, B. (2005). Opportunities for repeat testing: Practice doesn't always make perfect. *Applied H.R.M. Research*, *10*(2), 47–55. https://doi.org/10.1037/e518612013-432

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology*, *92*(2), 373–385. https://doi.org/10.1037/0021-9010.92.2.373

Kulik, J. A., Kulik, C.-l. C., & Bangert, R. L. (1984). Effects of practice on aptitude and achievement test scores. *American Educational Research Journal*, *21*(2), 435–447. https://doi.org/10.3102/00028312021002435

Lievens, F., Buyse, T., & Sackett, P. R. (2005). Retest effects in operational selection settings: Development and test of a framework. *Personnel Psychology*, *58*(4), 981–1007. https://doi.org/10.1111/j.1744-6570.2005.00713.x

Raymond, M. R., Neustel, S., & Anderson, D. (2007). Retest effects on identical and parallel forms in certification and licensure testing. *Personnel Psychology*, *60*(2), 367–396. https://doi.org/10.1111/j.1744-6570.2007.00077.x

Raymond, M. R., Neustel, S., & Anderson, D. (2009). Same-form retest effects on credentialing examinations. *Educational Measurement: Issues and Practice*, *28*(2), 19–27. https://doi.org/10.1111/j.1745-3992.2009.00144.x

Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th. ed.).

Suggested citation:

Zhou, J., & Cao, Y. (2020). *Does retest effect impact test performance of repeaters in different subgroups?* (Research Report No. RR-20-16). Educational Testing Service. https://doi.org/10.1002/ets2.12300

Action Editor: Sooyeon Kim

Reviewers: Skip Livingston and Longjuan Liang

ETS and the ETS logo are registered trademarks of Educational Testing Service (ETS). Other trademarks are the property of their respective owners.

Find other ETS-published reports by searching the ETS ReSEARCHER database at http://search.ets.org/researcher/