

Opportunities for Repeat Testing: Practice Doesn't Always Make Perfect

Allison M. Geving, Shannon Webb, and Bruce Davis
Psychological Services Incorporated

The effects of repeated testing opportunities on score gains were investigated using scores from a sample of real estate licensee candidates (N=9,226). Score gains were significant, but minimal. In addition, responding to the same items on multiple occasions did not aid score gains, but length of time between retakes did.

Although most research focusing on score changes over repeat test administrations has focused on academic testing (e.g., introductory psychology final exams, statistics unit exams, etc.) or ability testing (e.g., SAT or GRE), state licensing exams are also offered on a repeat basis and are subject to changes in scores over multiple administrations.

Thousands of individuals per year take state licensing exams to qualify for licenses in such areas as construction, real estate, insurance, and cosmetology. Real estate licensing, in particular, draws many candidates. For example, in 2003 in California alone, real estate licensing exams were administered over 100,000 times (California Department of Real Estate, 2004). For many state real estate licensing exams, candidates have several opportunities to retake the tests if they do not pass on the first attempt and, in fact, may retake the exam on multiple occasions within a short period of time.

Investigating the effects of these repeat test administrations on test scores is an important topic. Because the purpose of many licensing exams is to protect the public from harmful actions made by licensees, repeat testing opportunities may allow candidates to pass tests and obtain licenses without actually being highly knowledgeable (Millman, 1989). In other words, retesting opportunities may increase the risk of obtaining false positives, subverting the main objective of licensure testing.

However, even though repeat testing has often been considered problematic (Millman, 1989), little research has been conducted about how repeat testers perform over subsequent test administrations. This paper will focus on answering several questions about how test scores change for real estate licensing candidates. First, the actual mean change in test score over multiple administrations will be calculated. (Few past studies have looked at test score changes over more than two administrations). Second, a comparison in test score change will be made for tests with different amounts of repeated content to determine whether or not repeated exposure to the same items improves performance more than exposure to different (yet parallel) items. Lastly, this paper will explore the effect of time between

administrations to ascertain whether or not performance improves more or less if the time between administrations is longer.

Test Score Changes Over Repeat Administrations

Most studies focusing on academic testing of students and ability testing of new job applicants have revealed that test scores tend to improve for these test-takers over repeat test administrations. Friedman (1987) reviewed data obtained from 177 students enrolled in an introductory statistics course at a small liberal arts university. The average score gain from the initial test to the second test was 17 points, reflecting an increase of one letter grade for the exam. In addition, 92% of the repeat scores were higher when compared to the students' initial scores.

Juhler, Rech, From, and Brogan (1998) also found an increase in scores for undergraduates choosing to retake exams in an introductory algebra course. In this case, only students earning a B or less on an exam were given the option to retake the exam. Between 88% and 95% of examinees who retook an exam performed better on the exam the second time. With these exams, the test content was similar, but not identical, suggesting that the students mastered the material at a deeper level than merely regurgitating the same answers to the same questions. In addition, students did have the opportunity to study between test administrations.

Another study by Cates (2001) revealed that undergraduates in an educational psychology course were able to improve their test scores over repeat testing opportunities. Students had the opportunity to elect to retake up to four of the five course exams to improve their scores. (The repeat exams were parallel test forms equated on cognitive level, mean item discrimination, and mean difficulty level). In between the original tests and the retests, the instructor reviewed the original test answers and responded to students' questions about the test. Of the 202 retests taken to improve an original score, 139 (69%) represented improved performance and there was a mean gain of 3.5 percentage points. Thus, in situations where students were able to retake tests in their college courses, scores increased significantly on subsequent tests. However, this may have been due to the fact that the students were often provided with correct answers prior to each retest.

Similar test gains have been found over repeat administrations for employee selection tests. Hausknecht, Trevor, and Farr (2002) investigated score changes over identical administrations of cognitive ability and oral communication selection tests for 4,726 law enforcement job applicants. These candidates showed significant test improvements over the one-year gap between the first test and second test, possibly due to the fact that they were able to obtain feedback between test dates. In addition, they showed significant gains between the second test and third test. Interestingly, however, there were no significant differences between scores on tests 3 and 4. The authors suggested that test familiarity, reduced test anxiety, and increases in skill levels (due to feedback) resulted in increased scores for the first several administrations, but test practice effects subsequently eroded or disappeared.

Score gains were also shown for 66,303 recent college graduates taking the Professional and Administrative Career Examination (PACE) as a prerequisite for federal employment (Wing, 1980). These individuals took subtests in the following five areas: verbal, judgment, induction, deduction, and arithmetic. In addition, a

sixth subtest that was a parallel form of one of the previous tests was administered to determine if score gains occurred on that test due to practice effects. Small but consistent score increases from the first test form to the second were indeed shown. However, practice effects were least influential for items that tested general knowledge rather than logical reasoning.

Score changes over repeat testing opportunities have also been demonstrated with the Minnesota Multiphasic Personality Inventory (MMPI). Kelley, Jacobs, and Farr (1994) found that nuclear power plant employees who experienced regular MMPI screenings, provided more “normalized” profiles over repeat tests. The authors suggested that these employees became “test-wise” and were able to answer the questions in a way that made their profiles more “average”.

Thus, score increases have been demonstrated over repeat administrations of such employment screening tools as cognitive ability tests, communication tests, and the MMPI. These score improvements were attributed to increased test familiarity, decreased test anxiety, and increased skill levels. In accordance with the previous research pertaining to achievement tests (which are similar to licensure tests in that they test knowledge of specific content), it is hypothesized that test-takers in our study will also exhibit score increases over repeat testing (Hypothesis 1). It is expected that licensing candidates will exhibit these gains due to increased test familiarity, time to study the material further, and decreased test anxiety (due to test familiarity). Thus we also expect that there will be a positive correlation between the number of retakes and score change (Hypothesis 2a) and a positive correlation between the number of retakes and passing (Hypothesis 2b).

Effects of Repeated Content on Item and Test Performance

Several past studies have investigated how test scores change over repeat administrations when content is repeated from the original test to the next test. Kulik, Kulik, and Bangert (1984) meta-analyzed 40 studies investigating the effects of repeat testing opportunities on aptitude and achievement exam scores. They found that score gains were greater when the test forms were identical (effect size = .42, $k = 19$) than when they were parallel (effect size = .17, $k = 21$).

Krumboltz and Christal (1960) also found stronger practice effects for identical test forms than different tests in a study of score gains in aptitude testing. Air Force Reserve Officer Training Corps student officers took two spatial aptitude tests, seven hours apart. All received the identical form first and one-third received the identical test second, one-third received the parallel form second, and one-third received a completely different spatial relations test second. Similar positive effect sizes were found for the identical and parallel forms groups, but the group taking the different test second did not receive any gains from having taken the first test originally.

Similar results were found with the General Aptitude Test Battery (GATB) (United States Department of Labor, 1970), which includes the following subtests: intelligence, verbal, numerical, spatial perception, form perception, clerical perception, motor coordination, finger dexterity, and manual dexterity. Five-hundred-and-ten individuals took the GATB on a single test date and 156 of these individuals took the identical GATB again after two weeks, whereas 354 of these

individuals took a parallel test form after two weeks. Practice effects for the identical test were larger (effect sizes = .32 to .74) than for the parallel test (effect sizes = .15 to .55) for every test portion, with greater differences for the perceptual subtests than the intelligence, verbal, and numerical subtests.

Overall, research on the effects of repeat testing on score changes has shown that test-takers generally experience a greater score gain if they are taking an identical test the second time rather than an alternate form. Thus, it is expected that, in our study, score gain over repeat administrations will be greater for exams with more previously viewed items (Hypothesis 3a). It is also expected that candidates will be more likely to respond correctly to an item the second time if they already viewed the item on a previous test (Hypothesis 3b).

Amount of Time Between Retakes and Score Gains

Most of the research pertaining to time between retakes and score gains involves aptitude testing wherein candidates have taken psychomotor, logical reasoning, memory, and attitude inventories. Findings have ranged from no main effect for retest interval (Burke, 1997) to a decrease in practice effects as the interval increased (Caretta, Zelenski, & Ree, 2000).

However, the effects of time lapse between test administrations may be different for these types of exams than for licensure exams, which require individuals to learn and apply information. For licensure exams, individuals may gain more from studying test-related material between retakes than they would for reasoning or psychomotor tests. As very little research has been conducted on this topic, a hypothesis will not be offered at this time and these analyses will be exploratory.

Method

Participants

Participants were from one southern state who took a real estate salesperson licensing examination more than once between January 2, 2003 and July 31, 2004 and completed at least 10 items on each exam. In total, there were 9,226 individuals (gender information was not available).

Procedure

Participants took the licensing exam at various test centers located throughout the state. In total, these individuals took the examination 32,457 times over a seven-month period. Candidates were allowed to retake the exam as often as they wanted for six months before being required to file a new licensure application with the real estate commission.

The exam was computer-based and involved answering questions about property ownership, land use controls and regulations, valuation and market analysis, financing, laws of agency, mandated disclosures, contracts, transfer of property, practice of real estate, mathematics, specialty areas, commission duties and powers, licensing, standards of conduct, and agency/brokerage. The passing score for the exam was set at 70%. Each administration of the examination involved the presentation of 80 items from an item bank of 643 possible items. For each test,

items were selected to create tests with specific statistical properties (see Gibson & Weiner, 1998). Thus, the majority of items individuals received on each exam were different from items they viewed previously. The mean number of previously viewed items that a candidate received on each test was 9.81 out of 80 ($SD = 2.89$). Thus, on average, about 12% of the items were previously viewed. Test-takers were not provided with the correct answers between test administrations and were not allowed to review the test items at any point subsequent to each test administration. The only feedback provided was a description of the number of items that were answered correctly within each content area (e.g., property ownership, land use controls and regulations, etc.).

Results

Test Score Changes over Repeat Administrations

The mean number of retakes was 3.52 per person ($SD = 2.30$) and the mean number of days between retakes was 25.41 ($SD = 40.85$, range = 1 to 554). (See Table 1 for the mean scores for each test administration). The average change in score between the first test administration and the second test administration was 3.95 out of 80 ($SD = 6.38$) per exam, reflecting a significant [$t(9,225) = 59.51, p < .01, d = 0.62$] increase of 4.9%. This supported Hypothesis 1—that scores would improve over repeat administrations.

In regard to Hypothesis 2a, across all candidates and all re-takes there was a significant relationship between the number of retakes and the level of score gain, such that fewer re-takes resulted in a greater score gain between each administration of the test [$r = -.19, p < .001; F(25,23204) = 71.71, p < .001$]. Subsequent Tukey tests revealed significant differences between individuals who took the test two times and those who took it three to thirty-four times. Thus, hypothesis 2a was not supported.

Table 1
Overall mean score on each test administration

Variable	N	Mean	Std Dev
Test 1	9226	46.99	7.03
Test 2	9226	50.94	8.31
Test 3	5252	50.71	7.95
Test 4	3087	50.36	7.58
Test 5	1861	50.32	7.33
Test 6	1176	50.12	6.93
Test 7	775	49.19	7.49
Test 8	528	49.65	7.01
Test 9	358	49.14	6.82
Test 10	250	48.79	7.32

There was also a significant negative point-biserial correlation between number of retakes and passing [$r = -.22, p < .001$ ($n = 9,226$, number of retakes = 32,457)], revealing a lack of support for Hypothesis 2b. Tables 2 and 3 demonstrate that individuals who had extremely low scores on earlier tests often improved their scores over each retake, but not always enough to pass. (Note: individuals who had extremely low scores on earlier tests were more likely to retake the exam multiple times).

Effects of Repeated Content on Item and Test Performance

Hypothesis 3a, that the number of duplicate items on each exam would be related to points gained from the prior exam, was not supported ($r = -.004, p > .05, n = 9,226$), indicating that the test-takers did not receive any scoring benefits from having previously viewed items.

It was also hypothesized that test-takers would be more likely to respond correctly to a question the second time if they had previously viewed the item on another test. Across candidates who received the same items more than once, candidates responded with the same correct answer 50.30% of the time (Mean number of items = 4.96, $SD = 1.74$). Candidates responded with the same incorrect answer 26.57% of the time (mean number of items = 2.62, $SD = 1.36$). Candidates changed their answer to an incorrect one 14.14% of the time (mean number of items = .74, $SD = .14$), and changed to a correct answer 9% of the time (mean number of items = .63, $SD = .09$). Thus, on average, 64% of test-takers responded correctly to an item the first time, and 59% responded correctly to that same item the second time. This does not support Hypothesis 3b, in that candidates, on average, answered an item correctly about as often the second time they viewed the item.

Table 2
Mean score on each exam by number of times taken

Number of Retakes	Initial score	Score on Subsequent Exam									
		1	2	3	4	5	6	7	8	9	
1	49.12	55.42									
2	47.33	49.25	54.86								
3	45.87	48.06	49.28	54.21							
4	44.23	46.99	48.59	49.62	54.38						
5	43.15	45.90	47.84	48.40	49.74	54.15					
6	42.35	44.54	46.14	48.01	48.74	49.90	53.17				
7	41.45	44.21	45.32	46.60	47.94	48.94	48.95	53.48			
8	40.82	43.98	46.40	46.40	47.20	48.22	48.14	49.68	53.05		
9	39.10	41.30	42.29	43.25	44.62	45.80	45.87	47.04	47.45	48.79	

Note: values represent means only for candidates who took the exam X number of times. For example, the mean of 55.42 for retake #1 is based on scores from candidates who took the exam only 2 times

Effects of Time Between Test Administrations on Test Scores

Results showed that, as the number of days between test retakes increased, the score gain increased by a small but significant amount ($r = .02$, $p < .001$, number of retakes = 32,457).

Discussion

Test Score Changes over Repeat Administrations

The results of this study indicate that, on average, test scores do improve over repeat testing opportunities. A score increase of 4% can be practically and statistically significant, as the passing score was set at 70%, and, on the original test, 24% scored within 4% of the passing score. Previous authors have suggested that score gains could be due to several factors including: increased test familiarity, decreased test anxiety, and/or increased skill/knowledge. In this study, test-takers were provided with general feedback about the number of items answered correctly within each content area, but were not provided with a tutorial or item review. However, they did have the opportunity to study reference materials on their own in the time between retakes. This could have helped them increase their scores. In addition, they also gained familiarity with the structure of the test items, the test length, and time allotment. This increased familiarity could also have contributed to the score improvement. Lastly, it is possible that test anxiety was reduced after one administration due to an increase in test familiarity.

However, the results showed that the number of retakes and score gains were inversely related, indicating that, after the second testing opportunity, score gains were not as great. This result suggests that test-takers who must take the test more than two times may have trouble experiencing a large enough score gain to allow them to pass. Thus, repeat testing may not be that problematic in regard to allowing incompetent individuals opportunities to obtain licensure.

Table 3
Percentages of Passing Candidates (of the Total Number of Candidates Who Took the Test for the Designated Number of Retakes) on Each Retake

Test Administration	Percentage of Passing Candidates
Retake 1	31.6%
Retake 2	27.9%
Retake 3	22.8%
Retake 4	20.7%
Retake 5	18.0%
Retake 6	15.4%
Retake 7	17.6%
Retake 8	10.2%
Retake 9	13.8%

Note: Values are only for those candidates who took the exam X number of times. For example, the percentage that passed on retake #3 represents the percentage for only the candidates who tested three times.

Effects of Repeated Content on Item and Test Performance

It was hypothesized that candidates would be more likely to respond correctly to items they had viewed previously on another exam. Surprisingly, this was not the case. Previously viewing an item did not seem to affect performance on that item. This could have been due to several factors. First, test-takers were not provided with answer keys subsequent to any test administration nor were they allowed to review the items on their own. This may have made it difficult for them to learn how to answer these items correctly. In addition, the average number of days between retakes was 25. This may have been too long of a time period for them to recall the items when they were taking the test for a second, third, or fourth time.

Effects of Time Between Test Administrations on Test Scores

Larger retest intervals were associated with greater gains in test scores. It is likely that more days between test administrations boosted scores because individuals spent more time studying. This is supported by the fact that, of the 2,466 test-takers who ultimately passed, the time lapse between the exam prior to the passing exam was slightly more predictive of score change than for the overall group ($r = .13, p < .001$ vs. $r = .02, p < .001$).

Conclusions and Future Research Suggestions

Several conclusions can be drawn from this study. First, by retaking the real estate licensure exam, candidates may be able to show gains in test scores, but the largest gains in test scores generally occur over the first few test-taking opportunities. Second, including repeated content on multiple test administrations isn't as problematic for the licensing agency as expected. Candidates were not able to capitalize on the opportunity, as they were no more accurate in responding to these questions than in responding to new questions. Lastly, retest interval was positively correlated with test score gains, suggesting that candidates may want to wait longer between test administrations (and presumably study more) if they are interested in boosting their scores.

Future research should focus on several avenues of exploration. First, the factors that contribute most strongly to score improvements due to repeated testing should be investigated. It is not clear whether increased test familiarity, decreased test anxiety, or increased skill/knowledge contributes most strongly to test score gains. Second, candidates should be asked about what they do between testing opportunities. It would be useful to ascertain which study methods may be most helpful.

References

- Burke, E. F. (1997). A short note on the persistence of retest effects on aptitude scores. *Journal of Occupational and Organizational Psychology*, 70, 295-301.
- California Department of Real Estate. (2004). *Licensee Statistics for Fiscal Years 2002-2003 and 2003-2004*. Retrieved September 10, 2004 from <http://www.dre.ca.gov>.

- Carretta, T. R., Zelenski, W. E., & Ree, M. J. (2000). Basic Attributes Test (BAT) Retest Performance. *Military Psychology, 12*(3), 221-232.
- Cates, W. M. (2001). The efficacy of retesting in relation to improved test performance of college undergraduates. *Journal of Educational Research, 75*(4), 230-236.
- Friedman, H. (1987). Repeat examinations in introductory statistics courses. *Teaching of Psychology, 14*(1), 20-23.
- Gibson, W. M., & Weiner, J. A. (1998). Generating random parallel test forms using CTT in a computer-based environment. *Journal of Educational Measurement, 35*(4), 297-310.
- Hausknecht, J. P., Trevor, C. O., & Farr, J. L. (2002). Retaking ability tests in a selection setting: Implications for practice effects, training performance, and turnover. *Journal of Applied Psychology, 87*(2), 243-254.
- Juhler, S. M., Rech, J. F., From, J. F., & Brogan, M. M. (1998). The effect of optional retesting on college students' achievement in an individualized algebra course. *Journal of Experimental Education, 66*(2), 125-138.
- Kelley, P. L., Jacobs, R. R., & Farr, J. L. (1994). Effects of multiple administrations of the MMPI for employee screening. *Personnel Psychology, 47*, 575-591.
- Krumboltz, J. D. & Christal, R. E. (1960). Short-term practice effects on tests of spatial aptitude. *Personnel and Guidance Journal, 38*, 385-391.
- Kulik, J. A., Kulik, C. C., & Bangert, R. L. (1984). Effects of practice on aptitude and achievement test scores. *American Educational Research Journal, 21*(2), 435-447.
- Millman, J. (1989). If at first you don't succeed: Setting passing scores when more than one attempt is permitted. *Educational Researcher, 18*(6), 5-9.
- United States Department of Labor. (1970). *Manual for the USES General Aptitude Test Battery*. Washington, D.C.: U.S. Department of Labor.
- Wing, H. (1980). Practice effects with traditional mental test items. *Applied Psychological Measurement, 4*(2), 141-155.

Author Contact Information

Allison M. Geving, Ph.D.
PSI Licensure: Certification
2950 N. Hollywood Way, Suite 200
Burbank, CA 91505
1-800-367-1565
ageving@psionline.com