

Running Head: PRACTICE EFFECTS AND PROGRAM FEATURES

Practice Effects and Program Features in Professional Regulatory Examinations

Kirk Becker

Jim Masters

Jeannine Bailey

Pearson VUE, Evanston, Illinois

Paper presented at the annual
American Educational Research Association Meeting
Chicago, Illinois 2007

Practice Effects and Program Features in Professional Regulatory Examinations

Background

Candidates who retake examinations have been found to achieve higher scores. For candidates who retake an examination due to a previous failure attempt, their chance of passing may be greater simply because they have practice taking the examination, (Kulik, Kulik, & Bangert, 1984). While some gains in score may be attributable to candidate acquisition of knowledge and experience over time, gains may also occur because of previous experience with an examination, such as experience with the question format or examination length. These score gains, or practice effects, can give the examinee an unfair advantage, potentially passing individuals who may not have the necessary knowledge.

One way to limit unfair score gains is by constructing and administering multiple forms of a test. However, the development of new test forms requires substantial effort, with each step requiring additional time and cost. Additional item writing, review, and pretesting is required to build an item bank sufficiently large to support multiple test forms. Once a sufficient item bank is developed, the multiple forms must be constructed, equated, and published.

A number of test and candidate factors can influence score gains. Kulik et al. (1984) conducted a meta-analysis of 40 studies and found that examinees retaking a form identical to their initial form have larger score gains versus when a different but parallel form was used; this is in part due to a candidate's ability to remember specific items from the previous examination. Candidate ability also may affect score gains, with greater practice effects for higher ability candidates versus those with lower ability (Kulik et al., 1984). Using a real estate license examination, Geving, Webb, and Davis (2005) found that after the first retake, score gains

decreased with the repeated test administrations. It is hypothesized that with repeated test administrations, candidates become more familiar with examination properties, such as item format and test length, and test anxiety diminishes as candidates become accustomed to the examination and testing environment. This would aid candidates during their first retake examination, but would not be as beneficial for the remaining lower-ability candidates on subsequent retakes. (This may also be an artifact or interaction since candidates retaking multiple times are by definition lower ability than those who pass.)

Results from previous studies investigating test and candidate factors that influence score gains may not be applicable to licensure and certification testing (Geving et al., 2005; Anderson, Neustel, & Raymond, 2006). Most research conducted has involved aptitude testing, whereas licensure and certification examinations are more similar to achievement tests. For example, the Kulik et al. (1984) study only involved 7 achievement tests among the 40 studies analyzed. Aptitude and achievement tests are typically shorter than certification examinations (Anderson et al., 2006). Previous studies primarily involve school-aged children rather than adults. In addition, because only failing candidates retake credentialing examinations, the sample of retakers is composed of generally lower ability candidates. Research studies present further evidence of these differences. In a study involving the American Registry of Radiologic Technologists (ARRT), Anderson et al. (1996) found no evidence of score gains using identical versus parallel forms; a finding that differs from the Kulik et al. (1984) study.

Little research exists that investigates score gains among licensure and certification examinations (Geving et al., 2005). This study takes advantage of a large variety of archival testing program data to examine practice effects in professional licensure examinations across a range of test and candidate features. This archive of testing programs includes examinations from

many different programs across several professions. These examinations can vary not only in individual test and candidate features, but also in other program factors such as retake rules and candidate population. This study will investigate seven factors of testing programs that may affect score gains: three test-level features, three candidate-level features, and one program-level feature. Test-level features include test length, form difficulty, and number of forms used. Candidate-level features include time since last test administration, candidate ability, and number of retakes. The program-level feature included is total candidate volume. In addition to analyzing results across programs, this study will also compare results across programs within several professions. Breaking out results by profession, where possible, will help demonstrate whether relationships between practice effects and factors are variable across licensure programs or consistent. Because of the investigative nature of the study, the research is correlational rather than experimental.

Hypotheses are generated around the theory that more test experience, whether through individual experience or group sharing, will increase score gains due to practice effects. The amount of test experience can be determined by test features (such as increasing test length or available forms) or candidate features (such as number of retakes). A negative relationship is predicted for test-level features: there should be less of a score gain with increasing test length, increasing form difficulty, and increasing numbers of forms used. This is because the increase of these factors will decrease memory for specific items. Although Anderson et al. (1996) found no relationship between identical forms and score gains, this could be due to specific examination, candidate, or study properties (e.g., test length, candidate preparedness for the examination, or sample size) that may not generalize to licensure examinations.

In most cases, a positive relationship is predicted for both candidate-level features: For candidate-level features, score gains should increase with increasing time since last administration, replicating Geving et al. (2005), as the increase in experience and knowledge over time will outweigh the effects of diminishing memory for the test. There should also be an increasing score gain for higher ability examinees as they will be more likely to remember test information by linking test content to previous knowledge. Following Geving et al. (2005), candidates will have diminishing score gains with increasing number of retakes (negative relationship) because the lower ability candidates will be the candidates on the subsequent retakes; this diminishing candidate ability should outweigh the effects of test memory. A positive relationship is also predicated between the program feature and score gains; score gains should increase with candidate volume because more candidates will be able to see and share items.

Methods

The independent variables in this study include binomial and ordinal variables, and the purpose of the study is to identify features that may contribute to practice effects, rather than modeling score increases over time. Because of this, correlations will be computed between ordinal variables and score changes, while analysis of variances will be used to investigate the relationship between binomial variables and score changes. These analyses will be conducted within and across professions.

Data

Data from tests taken in 2006 were selected from 139 different licensure tests drawn from a larger overall sample of data. The sample only includes candidates who tested more than once in 2006 for a given test. The data include results from a range of professional licensure programs. For purposes of this study, the tests have been categorized in one of four distinct professions.

The sample sizes for each profession are shown in Table 1. The total sample size was 10,340. Profession 3 represented the largest number of tests taken (60.6%) followed by profession 4 (23.7%). Profession 1 & 2 represented the lowest number of tests taken (8.7% & 6.9% respectively).

Three test-level features were included in the analysis. The variable used for test length is the total number of scored items on a test. Form difficulty was estimated using the passing percentage of the test form in 2006 – a measure of test easiness. Because test easiness was used instead of test difficulty, the hypothesized relationship will be reversed. Number of forms is the number of parallel test forms available at a given time. Additionally, a binomial variable indicating whether a candidate was taking the same test form as last time was calculated. Figures 1 through 3 provide the distributions of scored items, test easiness, and number of test forms by profession. Descriptive statistics are presented in Table 2.

Three candidate-level features were included in the analysis. The days since the last test administration was calculated based on the difference between a test date and the last test date for the same test program. Candidate ability was estimated by the candidate's percent correct on a given test form. Number of retakes is the number of times a candidate tested for the same program in 2006. The distribution of these variables is provided in Figures 4 through 6.

Descriptive statistics are presented in Table 2.

The total candidate volume is the number of candidates who took any form of a test in 2006. The distribution of this variable is provided in Figure 7. Descriptive statistics are presented in Table 2.

The dependent variable of interest is the percent (%) change in score for a candidate. This variable was computed by calculating the difference between a candidate's score on a test and

his/her last score on the test. The difference was then divided by the number of items on the test to standardize the metric across tests with widely varying lengths. The distribution of score changes for each profession is provided in Figure 8. Descriptive statistics are presented in Table 3. Overall and within each profession there was an increase in percent (%) change, which represents a score gain from one test administration to the next.

Instruments

All data reflect performance on multiple choice items with four response options administered via computer.

Analyses

For binomial or nominal variables, ANOVAs were run with percent change as the dependent variable and the binomial or nominal variables as independent variables. For all other variables, correlations with percent change were calculated for the entire sample and separately for each profession. While this study did not look at interaction effects between independent variables, correlations between independent variables also were calculated to provide some indication of where interaction effects might be occurring.

Results

Results of the 4x1 ANOVA investigating the relationship between percent change and profession show a significant main effect of profession [$F(3, 10342) = 56.556, p < .001$].

Table 4 presents the correlations of the percent change with the seven different factors for each of the professions. The results in Table 4 show that some general trends can be seen across professions. For test features, a negative relationship was significant for profession 3 and a similar trend was seen for professions 1 & 2. The difference in sample sizes, with profession 3 having much higher sample size, could account for these differences. Profession 4, which had the

longest test length, showed no relationship. For test easiness, results support a significant positive relationship for the total group and for professions 1 and 3, and a non-significant positive relationship for profession 2. The negative relationship for profession 4 is non-significant. A negative relationship between number of test forms and score gain was supported for profession 3 and a general trend was seen for professions 1 and 4. Profession 2 only used 1 test form.

For candidate features, no significant correlation between score change and time since last administration was found. Professions 2 and 4 showed a positive trend, whereas profession 1 showed a negative trend. A positive relationship was supported for all 4 professions for candidate ability and score change. Finally a negative relationship was found for all professions and number of retakes.

Table 5 provides the comparison of score changes for candidates taking the same test form vs. parallel test forms on successive testing instances. Differences for the total sample and for separate professions were not significant at $p < 0.01$.

The analysis of potential interaction affects is not within the purview of this paper, however features of the testing programs analyzed could explain some of the results observed. For this reason, the correlations between the independent variables were generated and are shown in Table 6. This table may suggest areas where analyses of specific interactions would be informative.

Discussion

The results in Table 4 show that some general trends can be seen across professions. Overall and within each profession, the percent change in score from one test administration to the next was positive, representing a score gain. For test features, a negative relationship of score

gain with test length was predicted; this relationship was generally observed. A positive relationship for test easiness was predicted and observed for overall data, with professions either showing positive or non-significant correlations. A negative relationship between number of test forms and score gain was predicted, with decreasing score gain as number of forms increased. This hypothesis was supported as well. Table 5 supports Anderson et al.'s (2006) conclusion that there is no significant difference in score gains for candidates who retake the same form versus candidates who take a different form when they retest.

Overall, for test variables, all three of our hypotheses were supported, which includes a negative relationships for test length, test difficulty (or a positive relationship with test easiness), and # test forms. This supports the theory that more test experience will increase score gains due to practice effects. Practice effects can arise for many reasons. One such reason may be because with increasing practice, candidates better understand and remember test content and can use that information to better their performance. For example, test understanding and memory can be enhanced through test length where a candidate has a chance to experience an increased amount of items on an examination and thus obtain a better understanding of the examination and better remember test content in the future. Test difficulty may make the examination more difficult to predict and remember in the future, as well as more difficult to pass for the lower-ability candidates who need to retake an examination. Lastly, increasing the number of test forms may introduce variability of test content on subsequent retakes, making the examination more difficult to understand and remember. Cognitive theories of long-term memory and categorization may be useful for further explaining and predicting candidate performance in relation to test variables.

For candidate features, a positive relationship was predicted where score gains increase with increasing time since last administration, replicating Geving et al. (2005). Our results showed no significant correlation between score change and time since last administration. The negative trend observed for Profession 3 could be due to the increasing experience and studying acquired over time. A positive relationship was predicted between candidate ability and score gains, and this hypothesis was supported for all 4 professions. A negative relationship was predicted where candidates have diminishing score gains with increasing number of retakes, replicating Geving et al. (2005). This hypothesis was supported for all professions.

Overall, two of our three hypotheses for candidate features were supported, including the positive relationship for candidate ability and the negative relationship for number of retakes. Results suggest that candidate ability plays a larger factor than memory for tests. Candidates with higher ability may be able to better understand and gather more information from the examination to aid them on future examinations, but they are also likely to have increased score gains because they are higher ability candidates. Likewise, taking additional examinations may help a candidate further understand and remember the test, but those effects are out-weighed by the lower ability of the candidates who need to retake the subsequent tests and thus don't score as highly.

A positive relationship was predicated between the candidate volume program feature and score gains. We found a significant effect in the opposite direction for profession 3, but a positive trend for profession 2. It is unclear why there are opposing effects for professions. This could be because of other factors not measured in the study, such as communication and examination preparedness in specific fields.

The interaction of factors also may contribute to score gains. The investigation of interacting factors is beyond the scope of this study. However, correlations between factors were run to illuminate possible starting points for future investigation. Most correlations, although significant, were very weak. There were six correlations that showed slightly stronger correlations (six above $r = 0.20$ and four above $r = 0.30$). Several of those stronger relationships included test length; others included total tested. For example, there was a negative correlation between test length and test easiness, indicating that as examinations got longer, they also got more difficult. This could mean that score gains act differently across different combinations of these factors.

Factors contributing to score gains over time are complex and may affect different professions in varying ways. The examination of test, candidate, and program variables across several professions suggest that factors such as test memory and candidate ability contribute to score gains. However, the magnitude of these relationships is consistently rather small. The small magnitudes should be reassuring by helping to explain error variance in tests, but not bringing into question the validity of the examinations. Other factors that were beyond the scope of this study, as well as the interaction between variables, may further contribute to score gains. Other factors may include eligibility requirements, stakes of the test (required to work versus possible promotions versus prestige), and percent of item overlap from one form to another when there are multiple forms. Future research should be conducted to further understand these factors so that we may know when and how many additional forms need to be administered to prevent practice effects.

References

Anderson, D., Neustel, S., & Raymond, M. (2006, April). Item exposure and practice effects in professional certification: Do examinees benefit from seeing the same items on repeat administrations. Paper presented at the National Council on Measurement in Education Annual Meeting, San Francisco, CA.

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), pp. 47-56.

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), pp. 435-447.

Table 1.

Sample Size by Profession

Profession	Frequency	Percent
1	897	8.7
2	720	6.9
3	6271	60.6
4	2452	23.7
Total	10340	100

Table 2.

Mean, SD, and Min-Max for all Variables across and by Profession

Profession	All	1	2	3	4
	<i>Mean (SD) Min-Max</i>	<i>Mean (SD) Min-Max</i>	<i>Mean (SD) Min-Max</i>	<i>Mean (SD) Min-Max</i>	<i>Mean (SD) Min-Max</i>
Test Length	82.63 (30.7) 20-155	104.15 (12.82) 90-120	83.82 (7.89) 80-100	64.17 (29.45) 20-155	108.45 (19.49) 40-140
Test Easiness	0.5 (0.18) 0-0.93	0.34 (0.1) 0.24-0.59	0.64 (0.14) 0.33-0.88	0.56 (0.17) 0-0.93	0.49 (0.13) 0.15-0.8
Number of Forms	2.13 (0.86) 1-3	2.63 (0.48) 2-3	2 (0) 2-2	1.8 (0.94) 1-3	2.51 (0.77) 1-3
Days Since Last Test	30.48 (40.89) 0-308	47.74 (54.15) 2-266	28.43 (37.1) 3-217	25.27 (35.1) 0-247	24.08 (30.92) 1-308
Candidate Ability	0.62 (0.16) 0.12-0.98	0.74 (0.09) 0.34-0.94	0.6 (0.13) 0.28-0.83	0.6 (0.16) 0.12-0.98	0.55 (0.15) 0.15-0.82
Number of Retakes	3.1 (1.37) 2-14	2.63 (0.95) 2-9	3.64 (1.52) 2-9	3.16 (1.48) 2-14	3.25 (1.26) 2-11
Total Candidate Volume	1176.98 (1740.63) 3-7410	489.11 (305.24) 70-1042	727.55 (266.23) 109-1076	1437.57 (2262.75) 3-7410	1559.72 (1105.47) 64-3313

Table 3.

Mean, SD, and Min-Max for Percent (%) Change across and by Profession

Profession	All	1	2	3	4
	<i>Mean</i> <i>(SD)</i> <i>Min-Max</i>	<i>Mean</i> <i>(SD)</i> <i>Min-Max</i>	<i>Mean</i> <i>(SD)</i> <i>Min-Max</i>	<i>Mean</i> <i>(SD)</i> <i>Min-Max</i>	<i>Mean</i> <i>(SD)</i> <i>Min-Max</i>
Percent (%) Change	4% (8%) -19% to 48%	6% (7%) -16% to 0.27%	5% (7%) -15% to 30%	5% (8%) -19% to 48%	3% (6%) -18% to 25%

Table 4.

Correlations of Test, Content, and Program Features with Percent Change across and by Profession

	Total	Profession			
		1	2	3	4
Test Length	-0.09*	-0.06*	-0.09	-0.07*	0.01
Test Easiness	0.05*	0.13*	0.03	0.11*	-0.04
Number Forms	-0.08*	-0.03		-0.07*	-0.04
Days Since Last Test	0.01	-0.05	0.03	0.01	0.04
Candidate Ability	0.42*	0.5*	0.42*	0.44*	0.29*
Number of Retakes	-0.09*	-0.11*	-0.2*	-0.1*	-0.01
Total Tested	-0.03*	0	0.1*	-0.05*	0.02

* Significant at $p < 0.01$ (2-tailed).

Table 5.

Mean Percent (%) Change for Parallel vs. Same Test Form across and by Profession

	Total	Profession			
		1	2	3	4
Parallel Test Form	4.3%	6.1%	5.0%	4.7%	2.8%
Same Test Form	5.0%	6.6%	3.5%	5.1%	3.3%

Table 6.

Correlations of Test, Content, and Program Feature Variables

	Test Length	Test Easiness	Number Forms	Days Since Last Test	Candidate Ability	Times taken 2006	Total Tested
Test Length	1.000						
Test Easiness	-0.308*	1.000					
Number Forms	0.347*	-0.021	1.000				
Days Since Last Test	0.029*	-0.083*	-0.101*	1.000			
Candidate Ability	-0.080*	-0.073*	-0.148*	-0.005	1.000		
Number of Retakes	0.026*	-0.057*	0.020	-0.046*	-0.252*	1.000	
Total Tested	-0.338*	0.234*	0.366*	-0.062*	-0.129*	0.022	1.000

* Significant at $p < 0.01$ (2-tailed).

Figure 1.

Test Length (Number of Scored Items on Tests) by Profession by Profession

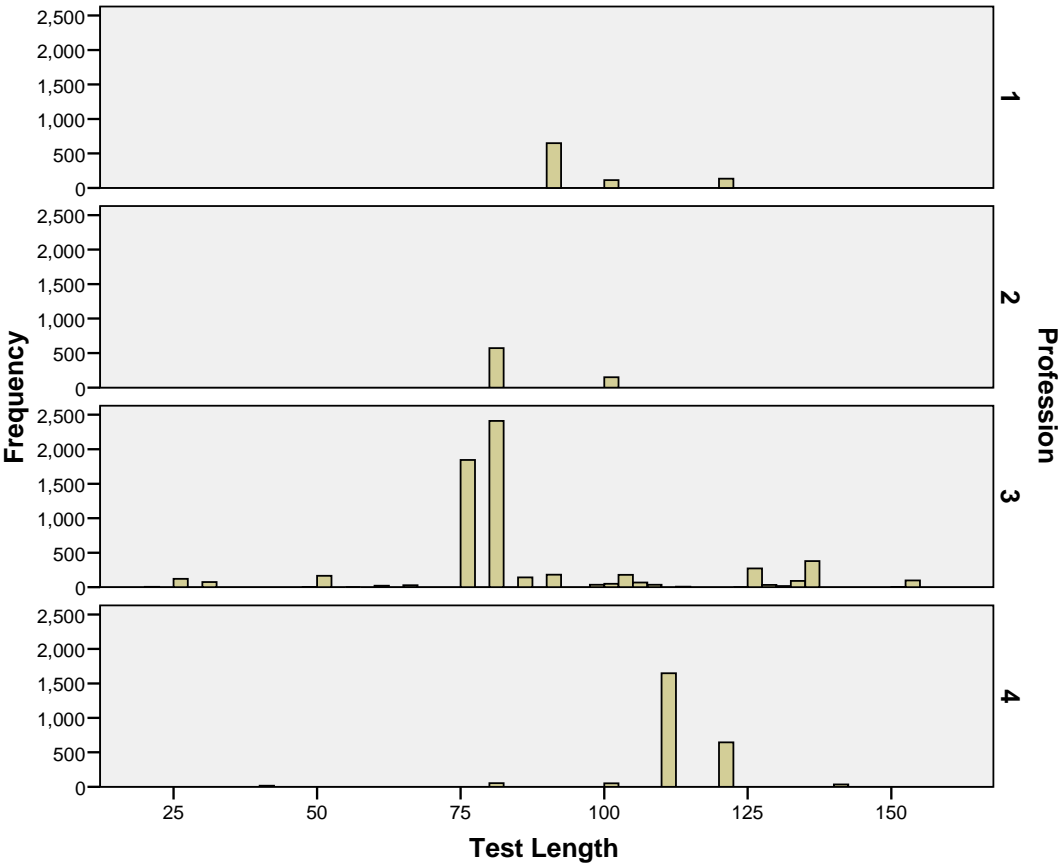


Figure 2.

Test Easiness (Passing Percentage (%) of Test Form) by Profession

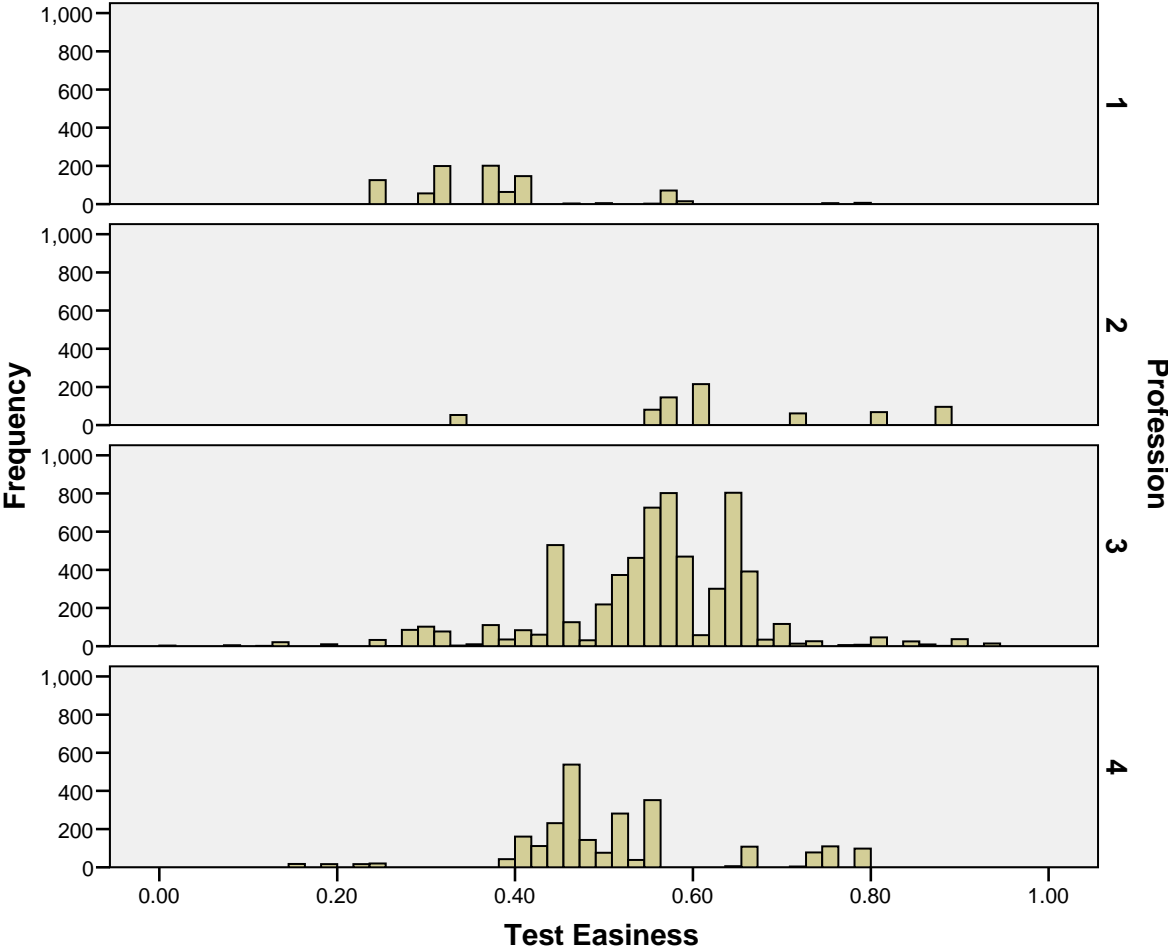


Figure 3.

Number of Test Forms by Profession

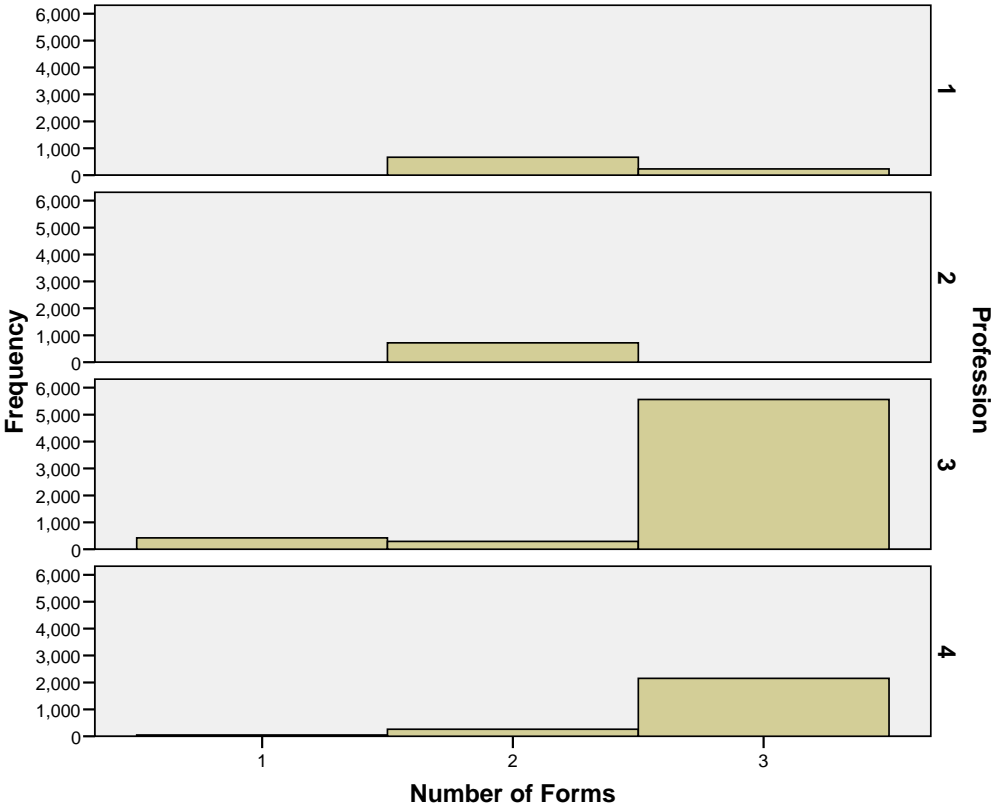


Figure 4.

Days since Last Test Administration by Profession

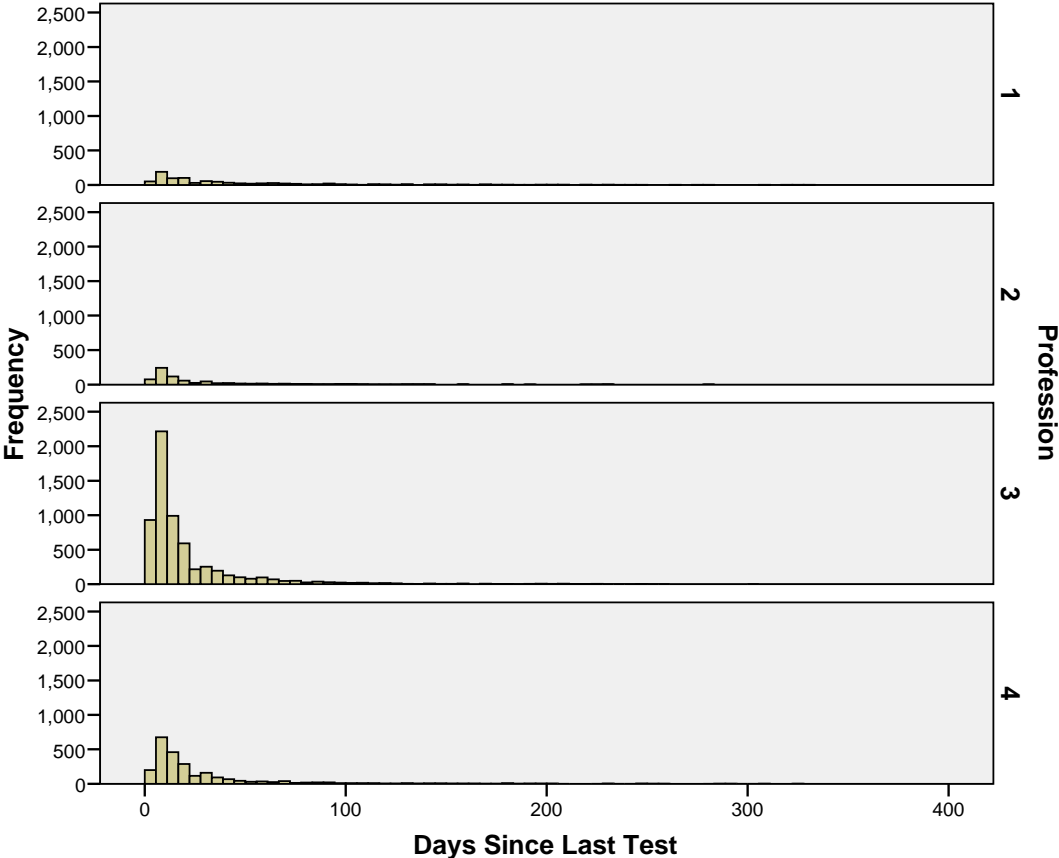


Figure 5.

Candidate Ability (Percent (%) Correct) by Profession

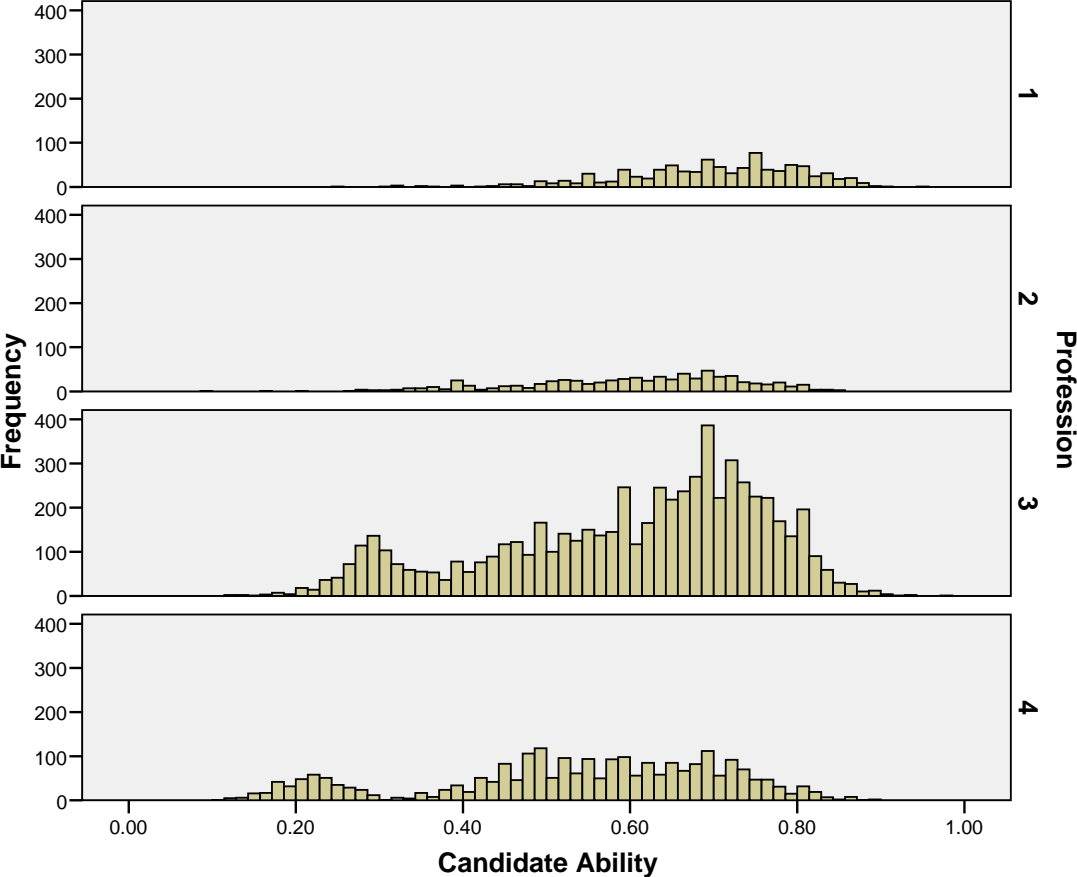


Figure 6.

Number of Retakes by Profession

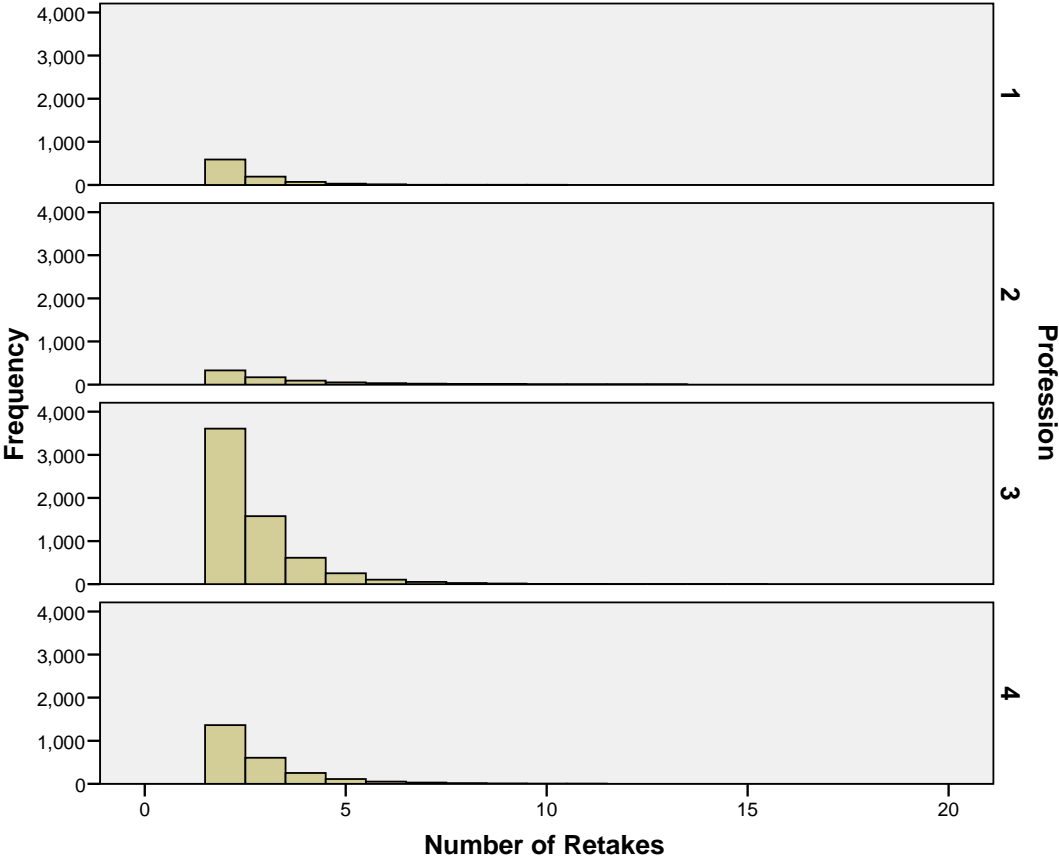


Figure 7.

Total Tested (Candidate Volume per Test) by Profession

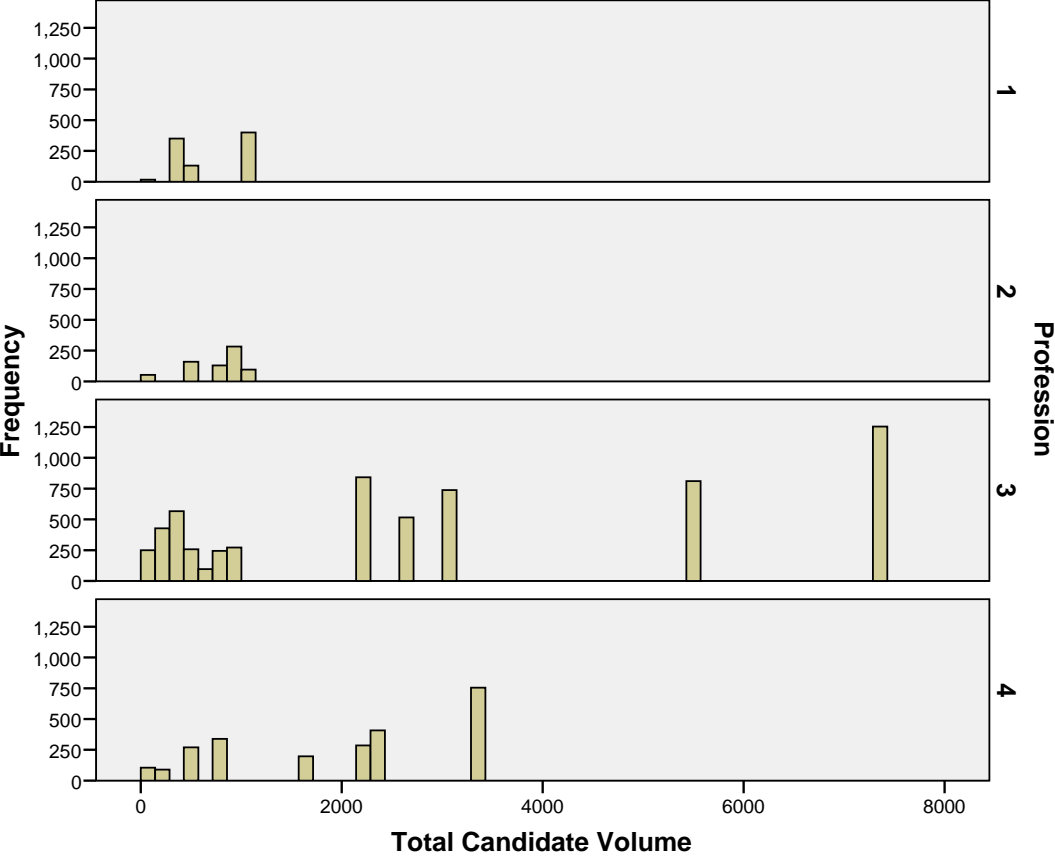


Figure 8.

Percent (%) Score Change by Profession

