ACT Research Report Series 90-1

Changes in PEP Test Scores for Students Who Have Tested More Than Once

Richard Sawyer
Catherine Welch

January 1990

ACT

For additional copies write:
ACT Research Report Series
P.O. Box 168
Iowa City, Iowa 52243

CHANGES IN PEP TEST SCORES FOR STUDENTS WHO HAVE TESTED MORE THAN ONCE

Richard Sawyer and Catherine Welch

ABSTRACT

The frequency of multiple testing on the Proficiency Examination Program (PEP) tests, the characteristics of examinees who retest, and the effect of retesting on test scores was investigated. Examinees who took a given PEP test more than once during the period from October, 1983 to October, 1987 were included in the study. Only examinees who failed on the first testing retested. Average increases between scores on the first and second testing were between 3.9 and 4.4 scale score points; however, some examinees have a negative gain on retesting. The results indicated only examinees whose first score was close to passing have a reasonable probability of passing on the second testing.

ACKNOWLEDGMENT

The authors thank Mark Houston for his help in the data analyses for this study.

CHANGES IN PEP TEST SCORES FOR STUDENTS WHO HAVE TESTED MORE THAN ONCE

The tests in the Proficiency Examination Program (PEP) are designed to measure subject matter primarily acquired outside the typical academic setting. They cover a broad range of academic disciplines and generally include material presented in a one- or two-semester course at the undergraduate level. Institutions use the test scores to award college credit, and to make decisions regarding course placement, course waivers, course requirements, and verification of skills required for certification. There are currently 42 tests in the PEP battery.

To maintain the security of the PEP tests, ACT does not release either the test booklets or the items after use. The multiple forms available for most PEP tests are administered according to a preplanned, nonreleased schedule. Examinees may retake a test only if 60 days have passed since they last took it. However, there is no limit to the number of times an examinee may take a test and the potential award of college credit is a significant incentive for examinees who fail a test to retake it.

The overall goals of this study are related to retesting on PEP tests. Specifically, we investigated: 1) the frequency of multiple testing on the PEP tests, 2) the characteristics of examinees who retest, and 3) the effect of retesting on test scores.

This research is relevant to counseling, because a crucial element in the decision of whether to retest is the anticipated increase in test score and the probability of passing the test. Many factors, such as additional experience in the subject areas tested, increased familiarity witk test format and testing procedures, and changes in the student's physical and emotional state, could be influential in this regard. No attempt is made here to differentiate among these possible factors or to test formal theoretical hypotheses about them. Rather, this study documents the distribution of observed test score changes for the overall group of students who elect to retake a PEP test. This information, when interpreted in the context of individual students' characteristics, should be helpful to students and counselors in deciding whether students should retake the test. Other appropriate information will, of course, also need to be considered in making decisions.

This research is concerned exclusively with observed scores and does not address issues of growth (i.e., increases in true scores). The assessment of growth is a very difficult and complex issue (Cronbach & Furby, 1970; Linn & Slinde, 1977), and a simple change score can be highly ambiguous for measuring growth. However, counselors may take into consideration the observed score change in conjunction with the processes taking place between two testings to try to explain the differences from one testing to a second testing (Gardner & Neufeld, 1987).

Method

Tests

For the purpose of this study, three high volume PEP nursing tests were selected. Each test is offered six times a year. Scores on all three tests are reported using a standard scale with a mean of 50, standard deviation of 10, and a range from 20 to 80. The recommended cut score for awarding credit is 45.

Descriptions of the three selected tests are given below. Each test contains between 140 and 155 multiple-choice items.

Commonalities in Nursing Care: Area A (427). Questions in the examination in Commonalities in Nursing Care: Area A are based upon common nursing problems and nursing care as they relate to the basic health needs of patients in the areas of safety, communications and interpersonal relations, comfort, rest and activity, maintenance of the integument, and asepsis. In addition, the health continuum and factors that affect health and illness are considered. Knowledge and understanding of technical vocabulary, anatomy, physiology, emotional and physical development, nutrition, and pharmacology are assumed.

Differences in Nursing Care: Area A (479). The Differences in Nursing Care examination measures knowledge and understanding relating to different health care problems frequently encountered by the associate degree nurse. Questions are based on the routine and specific manifestations of these problems and on the nursing care actions properly associated with them. Questions concern both acute and long-term problems of medical, surgical, psychiatric, obstetric, and pediatric patients. The examination requires the candidate to possess technical vocabulary and knowledge of anatomy and physiology, emotional and physical development, pharmacology, and nutrition generally expected of the associate degree nurse. The Differences in Nursing Care: Area A examination also considers nursing care of patients that is related to problems of oxygenation and problems of normal and abnormal cell growth.

Health Support: Area 1 (530). Questions focus on patterns that influence wellness and on the potential barriers to wellness. Emphasis is placed on the use of the nursing process to support the health of the client (individual, family, community) throughout the life cycle. Patterns of activity, sustenal patterns, developmental patterns, life-space, and the interrelationship of patterns are considered.

Data

Data for these investigations were obtained for examinees who took a given PEP test more than once during the period from October, 1983 to October,

1987. From the PEP history files for these four years, records were created containing the following variables: student name, mailing address, phone number, birthdate, sex, social security number (SSN), race, educational level, institution code, test center, test number, test form, test date, standard score, and item responses. Examinees who took a given PEP test more than once were identified from a variable formed by concatenating test number, student last name, and the first two characters of the first name. Records with identical values on the identification variable, and with identical values on either SSN or birthdate, were assumed to reflect actual repeated testings. All records not meeting these criteria were eliminated. A file consisting of one fixed length matched record per examinee, was then created for each test. The matched records contained all the data from repeated administrations of a given test. A separate file was created for each number of repeated testings, from two to ten.

The total test volume during the period October, 1983 to October, 1987 and the number of retested individuals, by test, are shown in Table 1.

Procedure

For each test, a frequency distribution of the number of testings was calculated, both for the total group and by racial/ethnic and sex group. Descriptive statistics on the scores obtained on the first testing, second testing, etc. were then calculated (when the number of observations on a given testing was greater than or equal to 25). Descriptive statistics on the scores obtained on the most recent (last) testing were also calculated.

Table 1

Test Volume and Number of Retested Individuals
October, 1983 to October, 1987

PEP test	Total test volume	Number of retested individuals
427	7628	906
479	6389	684
530	5453	468

The relationship between the test score obtained from the second testing and the test score obtained from the first testing was analyzed by linear regression analysis. Linear regression fits the "best" straight line to predict second test score from first test score. The line is "best" in the sense that the sum of the squared deviations between actual and predicted second test scores is a minimum (Draper & Smith, 1981).

The probability of obtaining a passing score on the second testing, given different scores on the first testing, was also estimated. Specifically, let the variable PASS equal 1 if the second test score is greater than or equal to 45 (the recommended passing score) and let PASS equal 0 if the second test score is less than 45. A regression of PASS on the first test score was conducted. Because the criterion variable is dichotomous, a nonlinear (logistic) regression model was used. A table of the resulting estimated probabilities of passing on the second testing, given the score on the first testing, was constructed.

Results

The relative frequencies of retesting observed for the period October 1983 to October 1987 for each of the three tests of interest are reported in Table 2.

Table 2

Relative Frequency of Retesting
(Proportion of Retesters)

Number of times	PEP Test		
tested	427	479	530
2	.74	.69	.77
3	.19	.19	.16
4	.04	.08	.05
5	.02	.02	.01
6	.01	.01	.01
7	.00+	.01	.01
8 or more	.00	.00+	.00+
(Number of records)	906	684	468

Of the 906 examinees who took Test 427 more than once, approximately 74% tested exactly twice and approximately 93% tested 2 or 3 times. Similar percentages hold for tests 479 and 530. Note that these percentages pertain to examinees who retested, not to all examinees.

Tables 3, 4, and 5 are crosstabulations of sex and race with number of times tested for Tests 427, 479, and 530, respectively. From Table 3, 17% of the retested examinees are Afro-American and 67% are Caucasian. Thus, for test 427, the proportion of Afro-American examinees tends to increase and the proportion of Caucasian examinees tends to decrease with the number of times tested. The other two tests show a more complex pattern. For all three tests, females tend to retest slightly more often than males.

Table 3

Proportions for Race and Sex Categories, by Number of Times Tested (Test 427)

Student group	Number of times tested							All retested students
	2	3	4	5	6	7	8	
Racial/ethnic								
Afro-American	.15	.21	.29	.25	.29	.33	.00	.17
American Indian	.02	.03	.00	.00	.00	.00	.00	.02
Caucasian	.68	.65	.54	.55	.71	.67	.00	.67
Mexican-American	.01	.00	.03	.05	.00	.00	.00	.01
Oriental	.02	.02	.06	.00	.00	.00	.00	.02
Spanish-speaking	.01	.00	.00	.00	.00	.00	.00	.01
Other/I choose not to respond/blank	.09	.09	.09	.15	.00	.00	.00	.09
Sex								
Female	.88	.88	.94	1.00	.86	1.00	.00	.89
Male	.12	.12	.06	.00	.14	.00	.00	.11
(Number of records)	671	170	35	20	7	3	0	906

Table 4

Proportions for Race and Sex Categories,
by Number of Times Tested
(Test 479)

	Number of times tested							
Student group	2	3	4	5	6	7	8	All retested students
Racial/ethnic								
Afro-American	.15	.17	.29	.17	.00	.25	.00	.17
American Indian	.01	.03	.02	.00	.00	.00	.00	.01
Caucasian	.66	.70	.56	.67	.56	.50	.67	.66
Mexican-American	.02	.01	.00	.00	.00	.00	.33	.02
Oriental	.02	.01	.00	.00	.00	.00	.00	.01
Spanish-speaking	.01	.01	.00	.00	.00	.00	.00	.01
Other/I choose not to respond/blank	.13	.08	.13	.17	.44	.25	.00	.12
Sex								
Female	.82	.84	.85	1.00	1.00	.75	1.00	.83
Male	.18	.16	.15	.00	.00	.25	.00	.17
(Number of records)	475	129	52	12	9	4	3	684

Table 5

Proportions for Race and Sex Categories, by Number of Times Tested
(Test 530)

	Number of times tested							
Student group	2	3	4	5	6	7	8	All retested students
Racial/ethnic								
Afro-American	.12	.20	.32	.00	.40	.67	.00	.15
American Indian	.01	.01	.05	.00	.00	.00	.00.	.01
Caucasian	.66	.58	.45	.60	.40	.00	.00	.63
Mexican-American	.03	.00	.05	.00	.20	.00	.00	.02
Oriental	.05	.05	.00	.00	.00	.00	1.00	.05
Spanish-speaking	.01	.05	.05	.20	.00	.00	.00	.02
Other/I choose not to respond/blank	.12	.09	.09	.20	.00	.33	.00	.13
Sex								
Female	.88	.89	.86	1.00	.90	1.00	1.00	.88
Male	.12	.11	.14	.00	.10	.00	.00	.12
(Number of records)	358	74	22	5	5	3	1	468

Means and standard deviations for each time tested (including the last time tested) are presented in Table 6 for each of the three PEP tests investigated. Results are reported when the number of observations for a given number of times tested was 25 or greater.

Table 6

Means, Standard Deviations (SD), and Sample Sizes (N), by Time Tested

	PEP Test								
	427			479			530		
Time tested	Mean	SD	N	Mean	SL	N	Mean	SD	N
First	37.1	6.4	906	38.1	5.9	684	39.6	5.1	468
Second	41.0	8.4	906	42.4	8.2	684	44.0	7.5	468
Third	40.6	8.1	234	42.3	7.0	209	43.0	6.8	110
Fourth	40.5	6.2	64	40.3	7.6	80	40.9	8.9	36
Fifth	38.8	8.3	29	40.8	7.6	28	---	---	---
Last	42.2	8.6	906	43.8	8.3	684	45.3	7.4	468

When interpreting these means, one should keep in mind that these statistics pertain to multiple-tested students. Note from Table 6 that the increases in means between the first and last time tested were 5.1, 5.7, and 5.7 standard score units for the PEP tests 427, 479, and 530, respectively.

Table 7 reports statistics associated with the linear regression of test scores obtained from the second testing on the test scores obtained from the first testing.

Table 7

Parameter Estimates for the Regression of Scores from the Second Testing on Scores from the First Testing

	Parameter			
Test	Intercept	Slope	r	RMSE
427	7.41	.91*	.70*	6.03
479	7.40	.92*	.66*	6.15
530	12.34	.80*	.54*	6.31

* $p < .001$

The intercept (a) and slope (b) in a regression analysis allow one to predict a student's second test score Y from his or her first test score X according to the formula $Y = a + bX$. The statistic r measures the consistency of the rank ordering across examinees for scores obtained on the first and second testing. RMSE (root mean-squared error) is an estimate of the variation of scores obtained on the second testing, given the score on the first testing; for examinees whose first test score is near the mean, about 2/3 will have a second test score within one RMSE of the predicted second score. For example, on test 427 the predicted score on the second testing for students with a score x on the first testing is $7.41 + .91x$, and about 2/3 of typical retesting students will have second test scores within 6.03 units of their predicted second test score.

Table 8 contains the estimated probabilities of passing a test on the second testing, given an examinee's score on the first testing, for the three PEP tests. Technical information on the logistic regression models employed are available from the authors, on request.

Table 8

Estimated Probabilities of Passing on the Second Testing, Given the Score on the First Testing

Score on the first testing	PEP Test		
	427	479	530
20	.00+	.01	.02
21	.01	.02	.02
22	.01	.02	.03
23	.01	.02	.04
24	.01	.03	.05
25	.01	.04	.06
26	.02	.04	.07
27	.02	.05	.08
28	.03	.07	.10
29	.04	.08	.11
30	.05	.10	.13
31	.07	.12	.16
32	.09	.14	.18
33	.11	.17	.21
34	.14	.20	.25
35	.17	.24	.28
36	.21	.28	.32
37	.26	.33	.36
38	.32	.38	.41
39	.38	.43	.45
40	.44	.48	.50
41	.50	.53	.55
42	.57	.58	.59
43	.63	.63	.64
44	.69	.68	.68

From Table 8, the estimated probabilities of passing on the second testing, given a score of 40 on the first testing, are .44, .48, and .50 for the PEP tests 427, 479, and 530, respectively. For tests 427 and 479, an examinee must have a score of at least 41 to have even or better than even odds of passing on the second testing. The probability of passing on the second testing, given a score of 30 on the first testing, are .05, .10, and .13.

Discussion

The results of this study pertain to average increases in observed scores on retesting. Only examinees who fail to obtain a 45 on the first testing retest. Therefore, these results only generalize to the average increase in observed scores for examinees who failed the first testing. The average increases between scores obtained on the first and second testing were 3.9, 4.3, and 4.4 for the PEP tests 427, 479, and 530, respectively. However, some examinees have negative gain on retesting (i.e., their scores decrease). Therefore, retesting does not guarantee a higher score. Further, only those students whose first score was close to passing have a reasonable probability of passing on the second testing. For students who score low on the first testing, the probability of passing on the second testing is quite small. Such students should probably not be encouraged to retest unless, of course, they have obtained additional experience or knowledge.

This research was concerned exclusively with observed scores and did not attempt to address issues of growth (i.e., changes in true scores). A measure of change, defined as the difference between the observed scores on two different testings, is very difficult to interpret. Given the acceptable reliability levels for these tests, a low correlation between scores obtained on the first and second testings may indicate that the test is not measuring the same traits on both occasions (Bereiter, 1963). Linn and Slinde (1977) stated "an item which measures problem-solving skill at one point in time may measure memory at a later point in time." (p.124). This may be the case with the PEP examinees who opt to retest. With moderate correlations between scores obtained on the first and second testings, some retesters may be gaining substantial experience and knowledge between the two testings, making the testing occasions very different from one another.

REFERENCES

Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. In C.W. Harris (Ed.), Problems in measuring change. Madison: University of Wisconsin Press, pp 3-20.

Cronbach, L. J. and Furby, L. (1970). How we should measure "change"--or should we? Psychological Bulletin, 74, 68-80.

Draper, N. and Smith, H. (1981). Applied regression analysis, second edition. New York: John Wiley.

Gardner, R. C. and Neufeld, R. W. (1987). Use of the simple change score in correlational analyses, Educational and Psychological Measurement, 47, 849-864.

Linn, R. L. and Slinde, J. A. (1977). The determination of the significance of change between pre- and posttesting periods. Review of Educational Research, 47, 121-150.