3-27-1981

The Effect of Multiple Retesting on Affect and Test Performance

Mary Lee Scott

Follow this and additional works at: https://scholarsarchive.byu.edu/dlls

BYU ScholarsArchive Citation
Scott, Mary Lee (1981) "The Effect of Multiple Retesting on Affect and Test Performance," Deseret Language and Linguistic Society Symposium: Vol. 7 : Iss. 1 , Article 5.
Available at: https://scholarsarchive.byu.edu/dlls/vol7/iss1/5

This Article is brought to you for free and open access by the All Journals at BYU ScholarsArchive. It has been accepted for inclusion in Deseret Language and Linguistic Society Symposium by an authorized editor of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.
THE EFFECT OF MULTIPLE RETESTING ON AFFECT AND TEST PERFORMANCE

Mary Lee Scott
Brigham Young University

The ostensible purpose of most tests, including language placement examinations, is to measure a student's knowledge and skill in a given subject area as accurately and as efficiently as possible. Unfortunately, this is not always as clear-cut a task as it may originally appear, as there seem to be several sources of variance in student scores in addition to a student's actual knowledge of the subject matter and random error of measurement. Many researchers (Millman, Bishop, and Ebel, 1965; Ford, 1973; Woodley, 1975; Wu and Slakter, 1978) have suggested that test wisdom is also a source of variance in student test scores.

Test wisdom or test "sophistication" (Erickson, 1972) has been defined by Millman, Bishop and Ebel (1965:707) as "a subject's capacity to utilize the characteristics and formats of the test and/or test taking situation to receive a high score. Test wisdom is logically independent of the examinee's knowledge of the subject matter for which the items are supposedly measures." Ebel and Damrin (1960) have suggested that test wisdom is a specific cognitive skill and as such, is capable of being developed through experience. It is felt to be responsible, at least in part, for the effects of practice and coaching on performance. In general, practice refers to actual experience in taking a certain type of test, such as a standardized achievement test or an IQ test, while coaching refers to training in test-taking techniques, including feedback on performance on sample test questions or alternate test forms (Berkeley and Sproule, 1973).

The results of the many studies examining the effects of practice and coaching on test performance are somewhat conflicting and inconclusive. In general, it seems that practice may result in a trend toward improvement in performance, although this improvement is not always statistically significant (Greene, 1937; Howard, 1964; Droege, 1966; Lane, 1966; Kreit, 1968; Mann, Taylor, Proger, Dungan and Tidley, 1970; Bowen, 1977). Other studies, examining the effect of coaching on test performance, have discovered that although an increase in test wisdom can be shown, these skills do not always generalize to an outside criterion test unrelated to the specific exercises used in training (Frankel, 1960; Wahlstrom, 1968; Lewis, 1971; Hecht, 1973; Woodley, 1975).

The effect of lack of test wisdom or test "sophistication" on performance (Berkeley and Sproule, 1973) is of genuine concern to test developers in the field of language testing as well. Several studies
have shown that the performance of students of similar ability but from different national backgrounds varies considerably on different test formats (Vernon, 1962; Millman and Setijadi, 1966; Lo and Slakter, 1973; Farhady, 1979). At most large universities and programs in English as a Second Language, applications for admission are received from students all over the world, and so their test wisdom or lack of it becomes an important issue in assessing the accuracy of the scores these students receive on entrance and placement examinations. At the same time, it becomes important to know which tests may cause the greatest amount of variance in scores due to practice effect.

Very little research has been done to date on the effect of practice in language testing. Bowen (1977) conducted a study in which 38 students at the American University in Cairo took five forms of the Michigan Test of English Language Proficiency (MTELP) over a period of ten weeks. Results revealed only small gains that were not statistically significant (about 0.7 points per administration), well within the standard error of measurement, across the test administrations. These findings led Bowen to conclude that there was no significant learning from practice. However, it should be noted that there was a regular tendency toward gain in score for the three subtests of the MTELP involved in the study (grammar, reading and vocabulary), although the gain was consistent only in the case of the reading test which showed gains with an interval of almost 25%. Bowen suggested that as the reading test was administered last and the three subtests were not separately timed, students may have learned to pace themselves better on later administrations, thus earning better scores. Bowen conceded "it could be argued that self-pacing and coping with mental fatigue are precisely the kind of elements of 'test wisdom' that practice effect is concerned with, and that therefore the reading test might well be the best evidence of practice effect." (p. 301). This conclusion is consistent with an outline of test wisdom strategies proposed by Millman, Bishop and Ebel (1965) which includes a time-using strategy including pacing and skipping or guessing at items which the examinee is unable to answer immediately.

With reference to another possible source of score variance, several researchers, as well as the American Psychological Association (1969), have suggested that lack of familiarity with test format or the testing situation may cause an examinee to experience anxiety which in turn may affect his or her performance on the test (Ammann, 1970; Berkeley and Sproule, 1973; Lange, 1978).

Generally appears that test anxiety and lack of test sophistication can influence the performance of examinees. As Berkeley and Sproule (1973:58) stated, "... individuals who are anxious about test taking or who are not sophisticated in test taking, perform less well on tests than they should. Their test scores do not accurately reflect their true levels of aptitude or achievement."
Methods

The purpose of this study was to examine the effects of multiple retesting on affect and performance on five different types of tests. The principle hypotheses examined in the study were that gain in performance due to practice effect would be different for different test types, and that student affect would vary significantly from test to test and for different administrations of the same test. In addition, it was hypothesized that Spanish-speaking students would differ significantly from Japanese students in their performance and affective reaction to different test types because of cultural and educational background.

The subjects in this study were students in the Brigham Young University Intensive English Program during the Spring Term of 1980. Seventy-three adult students participated in the study. Two major language backgrounds, Spanish and Japanese, were represented: sixty percent of the students were Spanish-speaking, while forty percent spoke Japanese. The range of student proficiency in English included beginning students, with little or no previous English instruction, to advanced. The students involved in the study were required to take the examinations discussed as part of an initial placement battery and subsequent evaluations during the term.

The students were divided into a high and low proficiency group based on their scores for a speaking test given on the day of the first test administration. Both the high and low proficiency students were administered a battery of three tests on three different occasions with approximately seven class periods between each administration. The content of each test was modified superficially each time in an attempt to minimize the effect of item recall from one administration to the next. Care was also taken to assure that testing conditions and procedures were as similar as possible for each administration to guard against a possible influence on the results of the study.

The speaking test was the same for both the high and low batteries and consisted of an oral interview lasting from eight to ten minutes, designed to assess a student's ability to communicate orally in English. In the course of the interview, opportunities were provided for the student to answer yes/no and information questions, respond to statements, and seek clarification. In addition to the oral interview, the high test battery consisted of a grammar test and a reading test. The grammar test was a modified version of the Integrative Grammar Test (IGT) (Bowen, 1975). In this test the students heard a series of sentences (spoken by an examiner) and were asked to write down the second word they heard in each sentence. A pause of ten seconds between each sentence allowed the students to record their answers. For example, the students might have heard a sentence such as, "Give 'm an inch and he'll take a mile." The students would record him as the second word they heard, even though the sentence was spoken at normal speed and only the m of him was actually heard. The reading test in the high battery involved an editing task. Thirty unnecessary and extraneous words had been added to a 220-word reading passage at the ninth grade reading level. These
words were chosen at random from the Dale list of 3,000 familiar words (Dale and Chall, 1948). The words were inserted after every third to fifteenth word in the passage using a table of random numbers. Students were allowed 10 minutes to locate the unnecessary words and cross them out.

As mentioned previously, the low test battery also included the oral interview, as well as a reading test and a listening test. The reading test required the students to identify as true or false statements referring to three picture series. The listening test was a modified version of the appropriate response section of the Alternate Modality Listening Exam (AMLEX) (Madsen, 1977). In this test students heard a question and three possible answers from a pre-recorded tape. The three possible answers were also written in their test booklets. The students were required to choose the best answer to the question on the tape.

After each administration of the high and low test batteries, students completed a Likert-style questionnaire designed to assess their affective reactions to the different test types, adapted from a similar questionnaire used by Jones, Madsen and Brown (1980). In this questionnaire students were required to evaluate each test on the basis of eleven criteria: 1) fairness, 2) how well the test corresponded to previous English instruction, 3) how well they liked the test, 4) how frustrating they found the test, 5) clarity of test instructions, 6) how well they felt they performed on the test, 7) how pleasant the experience of taking the test, 8) their perception of the difficulty of the test, 9) reliability, 10) validity, and 11) how well they felt the test reflected their knowledge of English. Responses to these criteria were assigned a numerical value from one to ten: one indicating a very negative response, and ten a very positive one. A final item on the questionnaire asked students to indicate whether or not they had ever taken a test exactly like any of the tests in the batteries administered in this study. The questionnaire was translated into Spanish and Japanese so that each student received it in his or her native language.

Results and Discussion

The tests administered in this experiment evidenced an acceptable degree of reliability. Calculations using the Kuder-Richardson formula 21 showed that the editing test in the high test battery had the highest reliability coefficient (.86), while the T/F picture test in the low battery had the lowest coefficient (.57). It was not possible to calculate reliability coefficients for the oral interview using the Kuder-Richardson formula 21 as the items on this test were not of equal difficulty, nor were they weighted equally when scored. Test-retest correlations were generally quite high in both batteries. The oral interview in the high battery correlated the highest with an average correlation of .96 (compared with .86 in the low battery), while the T/F picture test had the lowest average correlation, at .78.

Table 1, containing mean scores for the high and low batteries converted to percentages, shows a ranking of the tests in terms of
### TABLE 1

ACTUAL DIFFICULTY RANKING OF HIGH AND LOW
TEST BATTERY PERCENTAGE SCORES

<table>
<thead>
<tr>
<th>Test</th>
<th>High Battery</th>
<th>Low Battery</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(pts poss) (mean)(% of tot)</td>
<td>(pts poss) (mean)(% of tot)</td>
</tr>
<tr>
<td>IGT</td>
<td>50 25.88 56.76</td>
<td>Oral 50 21.61 43.22</td>
</tr>
<tr>
<td>Oral</td>
<td>50 35.20 70.40</td>
<td>AMLEX 40 20.53 51.33</td>
</tr>
<tr>
<td>Edit</td>
<td>30 22.57 75.23</td>
<td>T/F Pict 20 15.29 76.45</td>
</tr>
</tbody>
</table>

### TABLE 2

PERCENTAGE OF STUDENTS REPORTING
PREVIOUS EXPOSURE TO TESTS

<table>
<thead>
<tr>
<th>Test</th>
<th>Test</th>
<th>Japanese % tot</th>
<th>Spanish % tot</th>
<th>Total Students</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Oral</td>
<td>46.15</td>
<td>77.88</td>
<td>68.18</td>
</tr>
<tr>
<td></td>
<td>IGT</td>
<td>15.38</td>
<td>7.41</td>
<td>10.00</td>
</tr>
<tr>
<td></td>
<td>Edit</td>
<td>53.85</td>
<td>55.66</td>
<td>55.00</td>
</tr>
<tr>
<td></td>
<td>Oral</td>
<td>31.25</td>
<td>35.29</td>
<td>33.33</td>
</tr>
<tr>
<td></td>
<td>T/F Pict</td>
<td>12.50</td>
<td>29.41</td>
<td>21.21</td>
</tr>
<tr>
<td></td>
<td>AMLEX</td>
<td>43.75</td>
<td>23.53</td>
<td>33.33</td>
</tr>
</tbody>
</table>
their actual difficulty. A mean test score of 56.8% on the IGT indicates that it was the most difficult test in the high battery, while the oral interview (70.4%) was the next most difficult, and the editing test (75.2%) was the least difficult. As can be seen, there was quite a difference between the difficulty of the IGT and the other two tests which were fairly close in actual difficulty.

A mean score of 43.2% on the oral interview for the low proficiency students made it the most difficult test in the low battery. The AMLEX (51.3%) was the next most difficult test, and the T/F picture test was substantially easier than either the oral interview or the AMLEX.

As part of the affect questionnaire mentioned earlier, students were asked to indicate whether or not they had taken a test exactly like any of the tests administered in this study at any time prior to the study. Table 2 is a summary of those students who reported previous exposure to tests like those involved in this experiment.

In the high test battery, students reported being least familiar with tests similar to the IGT, more familiar with tests like the editing test, and most familiar with the format of the oral interview. The gap between the IGT and the other tests was considerable.

In the low test battery, students reported the least familiarity with the T/F picture format, and equal familiarity with tests similar to the oral interview and the AMLEX.

Both the Japanese and Spanish high proficiency students were least familiar with the format of the IGT, while the Japanese appeared most familiar with tests like the editing test, and the Spanish speakers with the oral interview. In fact, considerably more of the Spanish-speaking students were familiar with the oral interview than were the Japanese.

In the low test battery, the Japanese students were least familiar with the format of the T/F picture test, while the Spanish speakers were least familiar with that of the AMLEX. Again, the Spanish-speaking students were most familiar with the oral interview, while the Japanese reported more exposure to a test like the AMLEX.

The hypothesis that the gain in student performance attributable to practice effect would be different for different test types was generally supported by the results of the study. Results also tended to support previous research indicating that the greatest gain in performance resulting from practice effect seems to occur between the first and second test administration. In the high battery, the greatest increase in mean raw score between the first and second administration occurred on the IGT (6 points). This represented 12% of the total points possible. As will be recalled, this was the test format with which students had reported being least familiar. It seems logical to assume that an increase in performance occurred as students became more familiar with this format. The editing test showed almost as great a gain in raw score as the IGT (5.45 points),
which was actually a greater gain in percentage of total points possible (18%), between the first and second administration, even though students reported considerably more familiarity with it. The oral interview evidenced no significant gain (1% or .75 points). It should be noted that a greater percentage of the students reported being familiar with the oral interview format than with the other two tests.

The IGT showed a much smaller increase in mean student score (5% or 2.36 points) between the second and third administration, while the gain for the oral interview was larger (3% or 1.66 points) than that obtained between the first and second administration. This increase is still quite small when compared with student gains on the other two tests between the first and second administration. The mean score for the editing test remained virtually the same between the second and third administration.

In the low battery the oral interview showed the greatest gain (10% or 4.97 points) between the first and second administration. The AMLEX showed a gain of (8% or 3.16 points), and the T/F picture test an increase of (6% or 1.19 points). The T/F picture test showed a greater gain between the second and third administration (16% or 3.18 points), while the other two tests evidenced lesser gains (3 percent or 1.40 points for the oral interview and 2% or .85 points for the AMLEX). No clear relationship is apparent between performance gain in the low test battery and prior exposure to test format. An analysis of variance showed that the gains in performance across administrations were statistically significant only for the tests in the high battery.

Significant differences were observed in the performance of the Japanese and Spanish speaking students on the tests in both batteries, in support of another of the hypotheses of this study. This finding would seem to be related to results reported in the study by Farhady (1979) where students from different language and culture backgrounds performed at different levels on various types of tests. In the present study, the Japanese students scored higher than the Spanish-speaking students on all the tests in both batteries, although these differences were not statistically significant. This finding is surprising in view of the fact that the Spanish students generally reported being more familiar with the test formats than were the Japanese, with the exception of the IGT and the AMLEX.

An analysis of variance also demonstrated a significant difference in the affect ratings of the Japanese and Spanish-speaking students for the oral interview; the Spanish speakers rated it more positively than the Japanese on every item except their perception of the difficulty of the test, where their rating was quite similar to that of the Japanese students. This may have occurred at least partially because the Spanish were more familiar with the oral interview format than were the Japanese. It may also be due to a greater emphasis on speaking skills in the previous English instruction received by the Spanish-speaking students. The latter suggestion is merely tentative as this information was not available in this study.
The only affect rating for the oral interview which changed significantly across administrations was student evaluation of test difficulty. The students generally perceived the test as increasing in difficulty with each administration even though the results showed a trend toward improvement in their scores across administrations. Student perception of their performance remained fairly constant over the three administrations as well. Perhaps this test appeared less difficult to students at first because it was the least structured of all the tests. But as students became more familiar with it and compared experiences with their friends they may have perceived it as being more demanding than was originally apparent.

The hypothesis that student affect will vary significantly from test to test was strongly supported by the results of this study. A significant difference was revealed in the ratings for all of the tests in the high battery for all the affect items. For the tests in the low battery, affect ratings were significant for all items except test fairness, clarity of instructions, reliability, and how well the tests corresponded to previous English instruction.

In the high test battery, the oral interview was generally rated the most positively, the editing test the next most positively, and the IGI the least positively. It seems the students liked least the test that they were the least familiar with and which at the same time was the most difficult for them in terms of their actual performance.

In the low test battery, the mean affect ratings showed that students felt most positively about the T/F picture test, next most positively about the oral interview, and least positively about the AMLEX. Again, it will be recalled that the T/F picture test was the least difficult for the students, while the oral interview and AMLEX were considerably more difficult with regard to actual student performance.

Significant differences were also observed for the affect ratings of the two language groups for both test batteries, again supporting the original hypothesis that these groups would differ in their affective reaction to the tests. In general, it appeared that the ratings of the Spanish-speaking and Japanese students differed most for those items which involved more emotive reactions, such as how well they liked the tests, how frustrating they considered the experience of taking the tests, their perception of their performance, and how pleasant they considered the experience of taking the tests, the Spanish rating these items more positively than the Japanese. Other items seemed to involve more cognitive evaluations (how well the tests corresponded to previous instruction, clarity of instructions, test reliability, validity, and how well the tests reflected knowledge of English); and there did not generally seem to be as great a difference in the way the two language groups rated these items. It should be noted, though, that student evaluation of test fairness and difficulty seemed to involve both emotive and cognitive judgments.

In general, it was found that the Spanish-speaking students rated the tests in both batteries more positively than did the Japanese students. This trend included their evaluation of their performance,
even though as noted earlier, the Japanese had generally performed more successfully than the Spanish speakers on all of the tests. When this finding was reported at the Language Testing Conference at the University of New Mexico last summer, one of the participants in the conference who had lived for many years in Japan, suggested that this may have been a result of the Japanese culture, which requires that a person underestimate his performance or ability.

Although Japanese and Spanish-speaking students differed on several affect ratings, it is interesting to note that both groups felt the oral interview had the greatest validity and best reflected their knowledge of English.

Student affect was shown to vary for different administrations of the same test as well as among different test types. In the high test battery, students felt the tests were significantly less frustrating on the second and third administration and the instructions clearer. They also felt their performance had improved on the latter two administrations. In the low test battery, students considered the tests less frustrating with each administration. In general, it would seem that these findings support earlier conclusions by other researchers (Sassenrath, 1967; the American Psychological Association, 1969; Lange, 1978) that greater familiarity with test format leads to a decrease in the anxiety aroused by the testing situation.

Conclusions and Recommendations

The results of this study generally support the hypotheses originally postulated. Significant gains in performance were revealed for the tests in the high battery, while a general trend toward an increase in score was demonstrated for the tests in the low battery, although this trend was not statistically significant. The greatest improvement in test performance was generally found to occur between the first and second administration of the tests.

In conclusion, it is recommended that further study of the effect of practice on test performance and affect be undertaken in an effort to improve and refine current strategies in language testing. Research should be extended to include other language backgrounds besides Spanish and Japanese in order for findings such as those encountered in this study to be generalizable to many different language testing needs. In the meantime, as Farhady (1979) has suggested, perhaps batteries should be composed of various formats to avoid being unfair to one cultural background or another.

Finally, it appears important to begin to implement the findings of this and other similar studies by selecting language tests for use in placement batteries or for other purposes which are less susceptible to the effects of practice and negative test affect, while still accurately assessing a student's proficiency in the target language.
References

American Psychological Association, Task Force on Employment Testing of Minority Groups

Amman, Rene

Berkeley, Charles, S. and Sproule, Charles F.
1973 "Test anxiety and test unsophistication; the effects, the cures." Public Personnel Management 2 (January):55-59.

Bowen, J. Donald

Droege, R.

Ebel, R. L., and Damrin, D. E.

Erickson, M.

Farhady, Hossein

Ford, Valerie A.
1973 Everything You Wanted to Know About Test-Wiseness. Summer Graduate Student, Educational Testing Service.

Frankel, E.

Green, E. B.
Hecht, James T.

Howard, K. I.
1964 "Differentiation of individuals as a function of repeated testing." Educational and Psychological Measurement 24:875-894.

Kreit, L. H.

Lane, R. G., Penn, N. E., and Fischer, R. F.

Lange, Robert
1978 "Flipping the coin from test anxiety to test wisdom." Journal of Reading 22 (December): 274-277.

Lewis, Ernest L.

Lo, M. I., and Slakter, M. J.

Mann, Lester, Taylor, Raymond G., Jr., Proger, Barton B., Dungan, Roy H., and Tidey, William J.

Millman, J., Bishop, C., and Ebel, R.

Millman, J. and Setijadi
Vernon, Philip E.
1962 "The determinants of reading comprehension."
Educational and Psychological Measurement 22:
269-86.

Wahlstrom, M. and Boersma, F.
1968 "The influence of test wisdom upon achievement."
Educational and Psychological Measurement 28:
413-420.

Woodley, Katheryn K.
at the annual meeting of the National Council on
Measurement in Education (Washington, D.C., March 31-
April 2.)

Wu, Tieh-Hsiung and Slakter, Malcolm J.
1978 "Risk taking and test wisdom by Chinese students
by grade level and residence area." Journal of